Arbeitsstelle Interkulturelle Konflikte und gesellschaftliche Integration (AKI) (ed.)

The Effectiveness of Bilingual School Programs for Immigrant Children

Bearbeitung: Janina Söhn

Best.-Nr./Order No.: **SP IV 2005-601**

Wissenschaftszentrum Berlin für Sozialforschung (WZB)

Veröffentlichung der Arbeitsstelle Interkulturelle Konflikte und gesellschaftliche Integration (AKI) – *Programme on Intercultural Conflicts and Societal Integration (AKI)*

Juni 2005

Contents

Introduction

Is bilingual education a promising and effective type of instruction conducive to enhancing the academic achievement of immigrant children? This question was the subject of the workshop "The Effectiveness of Bilingual School Programs for Immigrant Children" organized by the Programme on Intercultural Conflicts and Societal Integration (AKI) at the Social Science Research Center Berlin (WZB). Experts from the United States, the Netherlands, Sweden and Germany gathered in November 2004 to discuss respective national experiences and evidence on the effects of bilingual education. This volume presents the revised papers.

The conference formed part of our Programme's (AKI) work on the broader topic of language and the incorporation of immigrants. The specific issue of bilingual education highlights two key problems linked to the process of integration: the representation of immigrants' culture in the education system and the challenge of overcoming unequal opportunities in education. In a separate research review (AKI-Forschungsbilanz „Zweisprachiger Schulunterricht für Migrantenkinder. Ergebnisse der Evaluationsforschung zu seinen Auswirkungen auf Zweitspracherwerb und Schulerfolg" by J. Söhn) we present our own assessment of the evidence available in evaluation studies and previous meta-analyses on the effectiveness of bilingual education.

In general terms, 'bilingual education' refers to models of school education where instruction is given in two languages and content matter is taught in two languages as well. In elementary education, the most important characteristic of bilingual instruction is that students are taught to read and write in both their native tongue and a second language – either starting at about the same time (simultaneous alphabetization) or consecutively (transitional models). There are further variants of bilingual education: The scope of teaching given in the immigrant children's first language can vary considerably both in terms of teaching hours per week and the number of years of instruction in this language. In some programs additive bilingualism is the pedagogical goal, while in others teaching the first language only serves as a means of promoting success in mainstream classes. In some national contexts, it is more common to use the term 'mother tongue' or 'home language teaching', rather than bilingual education. As our focus is on immigrants – either first or second generation – we are mainly concerned with cases in which children are taught both in the official language of the host country and in their dominant first language.

In all four countries represented in this volume, bilingual education or mother tongue teaching tends to be politically contentious – especially in the US where after fierce political debates a majority of citizens in California and Arizona voted for its abolition. Bilingual education is often taken as a symbol of how a plural society treats or should treat cultural and linguistic minorities. It can stand for the idea of embracing and strengthening multilingual diversity, whereas opponents of bilingual education see it as contradicting their prior aim of furthering social coherence by way of a common lanugage. Scientific evidence has

played a prominent role in the political struggles around this issue. However, research is not unanimous about the effects of bilingual education. Proponents and opponents in these debates draw on contrasting research results in order to support their normative arguments.

One major difference between the US and the three selected European countries – the Netherlands, Sweden and Germany – is the amount of research on this issue, i.e. of studies which evaluated bilingual school programs for immigrant children in the respective national context.

The following chapters by Robert E. Slavin/Alan Cheung and by Christine E. Rossell/Julia Kuder represent competing efforts to clarify which of the numerous studies on the US experience have a methodologically sound basis and what overall conclusions should be drawn: Is bilingual education more beneficial for the acquisition of a second language and academic achievement than monolingual instruction? And which type of bilingual model seems more promising? Slavin and Cheung present a (revised) meta-analysis of existing evaluation studies in the US, which are mainly English-Spanish models. They criticize an early attempt by Rossell to review this field of research. Rossell and Kuder, in turn, reanalyzed the meta-analysis by Slavin and Cheung and alternative meta-analyses. While Slavin and Cheung conclude that learning to read and write simultaneously in one's first and a second language is to be recommended, Rossell and Kuder do not fully object to this model of bilingual education, yet argue strongly against transitional bilingual education, which has been widespread and at times obligatory for English language learners in the US.

Geert Driessen and Monica Axelsson survey the history of bilingual education as an educational policy for immigrants in the Netherlands and Sweden respectively. The models of mother tongue teaching practiced in these countries are rather weak models of bilingual education because content matter is only rarely taught in the minority language. Yet teaching the children's native tongues as such has been widespread. Furthermore, the arguments in political debates put forward in favor of and against mother tongue teaching resemble those brought up in controversies about strong models of bilingual education. Both authors present research results on the effects of bilingual education and its correlation with the students' school careers in general. What is striking about the Dutch and the Swedish case is that after general support for mother tongue teaching in both countries in previous decades, governments have made different decisions in the new millennium: Sweden still supports mother tongue teaching, while the Netherlands abolished it altogether in August 2004.

In Germany, bilingual models in which instruction is given in German and an immigrant-minority language are still very rare and evaluation studies are almost non-existent. Ingrid Gogolin gives a brief overview of these studies. As she points out in her article, the German education system gives insufficient support to students whose native tongue is not German – with regard to both bilingual programs and other special measures for immigrant students. Complementary mother tongue instruction, for instance, has been offered in several regional states (Länder), yet under pedagogically unfavorable circumstances. It is thus unsuitable for an assessment of the effects of bilingual education. The development of

programs and teaching methods for children who grow up bilingually as well as empirical research on the effectiveness of such measures is urgently needed.

Future evaluation research about bilingual education models in Germany and other programs targeting German language learners will require new methodological instruments for the assessment of a bilingual child's command of German and her or his first language. Evaluation studies should be able to measure the level of language proficiency both at different points in time and across groups. Hans H. Reich presents preliminary results of an on-going project on the longitudinal assessment of bilingualism in the context of which he and colleagues have developed and applied these kinds of measurement instruments. In his contribution on the bilingual language development of German-Turkish children in primary schools, he addresses the issues of interdependence, i.e. interconnectedness and transfer of competences between first and second language, which lies at the heart of many theoretical arguments in favor of and against bilingual education programs.

Together the chapters in this publication offer a summary of the US debate on the effectiveness of bilingual education – represented by major participants – and give an overview of the experience and research in three European countries. By publishing this collection we hope to contribute to a so far underdeveloped international debate.

Janina Söhn, AKI, June 2005

A Synthesis of Research on Language of Reading Instruction for English Language Learners[1]

Robert E. Slavin and Alan Cheung

The reading education of English language learners (ELLs) has become one of the most important issues in all of educational policy and practice. As the pace of immigration to the U.S. and other developed countries has accelerated in recent decades, increasing numbers of children in U.S. schools come from homes in which English is not the primary language spoken. As of 1999, 14 million Americans ages 5-24, or 17% of this age group, spoke a language other than English at home. This is more than twice the number of such individuals in 1979, when only 9% of Americans ages 5-24 spoke a language other than English at home (NCES 2004). While many children of immigrant families succeed in reading, too many do not. In particular, Latino and Caribbean children are disproportionately likely to perform poorly in reading and in school. As No Child Left Behind and other federal and state policies begin to demand success for all subgroups of children, the reading achievement of English language learners is taking on even more importance. Thousands of schools cannot meet their adequate yearly progress goals, for example, unless their English language learners are doing well in reading. More importantly, American society cannot achieve equal opportunity for all if its schools do not succeed with the children of immigrants.

Sixty-five percent of non-English speaking immigrants in the U.S. are of Hispanic origin (NCES 2004), and this is also one of the fastest growing of all groups. Hispanics have recently surpassed African Americans as the largest minority group in the U.S. Hispanic students as a whole, including English proficient children in the second generation and beyond, score significantly lower in reading than other students. On the National Assessment of Educational Progress (NAEP; Grigg, Daane, Jin, & Campbell 2003), which excludes children with the lowest levels of English proficiency from testing, only 44% of Latino fourth graders scored at or above the "basic" level, in comparison to 75% of Anglo students. Only 15% of Latino fourth graders scored at "proficient" or better compared to 41% of Anglos. Further, 31% of students who speak Spanish at home fail to complete high school, compared to 10% of students who speak only English (NCES 2004).

There is considerable controversy, among policymakers, researchers, and educators, about how best to ensure the reading success of English language learners. While there are many aspects of instruction that are important in the reading success of English language learners, one question has dominated all others: What is the appropriate role of the native lan-

[1] Adapted from Slavin & Cheung (in press), A synthesis of research on language of reading instruction for English language learners, Review of Educational Research.

guage in the instruction of English language learners? In the 1970s and 1980s, policies and practice favored bilingual education, in which children were taught partially or entirely in their native language, and then transitioned at some point during the elementary grades to English-only instruction. Such programs are still widespread, but from the 1990s to the present, the political tide has turned against bilingual education, and California, Arizona, Massachusetts, and other states have enacted policies to greatly curtail bilingual education. Recent federal policies are restricting the amount of time children can be taught in their native language. Among researchers, the debate between advocates of bilingual and English-only reading instruction has been fierce, and ideology has often trumped evidence on both sides of the debate (Hakuta, Butler, & Witt 2000).

This article reviews research on the language of reading instruction for English language learners in an attempt to apply consistent, well-justified standards of evidence to draw conclusions about the role of native language in reading instruction for these children. The review applies a technique called "best-evidence synthesis" (Slavin 1986), which attempts to use consistent, clear standards to identify unbiased, meaningful information from experimental studies and then discusses each qualifying study, computing effect sizes but also describing the context, design, and findings of each study. Best-evidence synthesis closely resembles meta-analysis, but it requires more extensive description of key studies. Details of this procedure are described below. The purpose of this review is to examine the evidence on language of instruction in reading programs for English language learners to discover how much of a scientific basis there is for competing claims about effects of bilingual as opposed to English-only programs, in order to inform practitioners, policymakers, and researchers about the current state of the evidence on this topic as well as gaps in the knowledge base in need of further scientific investigation.

Language of Instruction

For many years, researchers, educators, and policy makers have debated the question of the appropriate language of reading instruction for children who speak languages other than English. Proponents of bilingual instruction argue that while children are learning to speak English, they should be taught to read in their native language first, to avoid the failure experience that is likely if children are asked to learn both oral English and English reading at the same time. Programs based on this philosophy transition children to English-only instruction when their English is sufficient to ensure success, usually in third or fourth grade. Alternatively, many bilingual programs teach young children to read both in their native language and in English at different times of the day or on alternating days. There is a great deal of evidence that children's reading proficiency in their native language is a strong predictor of their ultimate English reading performance (Garcia 2000; Lee & Schallert 1997; Reese, Garnier, Gallimore, & Goldenberg 2000), and that bilingualism itself does not interfere with performance in either language (Yeung, Marsh, & Suliman 2000). Bilingual advocates also argue that without native language instruction, English language learners are likely to lose their native language proficiency, or fail to learn to read in their native language, losing skills that are of economic and social value in the world today. Op-

ponents of bilingual education, on the other hand, argue that native language instruction interferes with or delays English language development, and relegates children who receive such instruction to a second-class, separate status within the school and, ultimately, within society. They reason that more time on English reading should translate into more learning (see Rossell & Baker 1996).

Reviews of the educational outcomes of native language instruction have reached sharply conflicting conclusions. In a meta-analysis, Willig (1985) concluded that bilingual education was more effective than English-only instruction. Wong-Fillmore & Valadez (1986) came to the same conclusion. However, a review by Rossell & Baker (1996) claimed that most methodologically adequate studies found bilingual education to be no more effective than English-only programs. Greene (1997) re-analyzed the studies cited by Rossell & Baker and reported that many of the studies they cited lacked control groups, mischaracterized the treatments, or had other serious methodological flaws. Among the studies that met an acceptable standard of methodological adequacy, including all of the studies using random assignment to conditions, Greene found that the evidence favored programs that made significant use of native language instruction. August & Hakuta (1997) concluded that while research generally favored bilingual approaches, the nature of the methods used and the populations to which they were applied were more important than the language of instruction per se. Quantitative research on the outcomes of bilingual education has diminished in recent years, but policy and practice are still being influenced by conflicting interpretations of research on this topic. The following sections systematically examine this evidence to attempt to discover what we can learn from research to guide policies in this controversial arena.

English Immersion and Bilingual Programs

When a child enters kindergarten or first grade with limited proficiency in English, the school faces a serious dilemma. How can the child be expected to learn the skills and content taught in the early grades while he or she is learning English? There may be many solutions, but two fundamental categories of solutions have predominated: English immersion and bilingual education.

English Immersion: In immersion strategies, English language learners are expected to learn in English from the beginning, and their native language plays little or no role in daily reading lessons. Formal or informal support is likely to be given to ELLs to help them cope in an all-English classroom. This might or might not include help from a bilingual aide who provides occasional translation or explanation, a separate English as a Second Language class to help build oral English skills, or use of a careful progression from simplified English to full English as children's skills grow. Teachers of English language learners might use language development strategies, such as total physical response (acting out words) and realia (concrete objects to represent words), to help them internalize new vocabulary. They might simplify their language and teach specific vocabulary likely to be unfamiliar to ELLs (see Calderón 2001; Carlo et al. 2004). Immersion may involve placing

English language learners immediately in classes containing English monolingual children, or it may involve a separate class of ELLs for some time until children are ready to be mainstreamed. These variations may well have importance in the outcomes of immersion strategies, but their key common feature is the exclusive use of English texts, with instruction overwhelmingly or entirely in English.

Many authors have made distinctions among different forms of immersion. One term often encountered is "submersion," primarily used pejoratively to refer to "sink or swim" strategies in which no special provision is made for the needs of English language learners. This is contrasted with "structured English immersion," which refers to a well-planned, gradual phase-in of English instruction relying initially on simplification and vocabulary-building strategies. In practice, immersion strategies are rarely pure types, and in studies of bilingual education, immersion strategies are rarely described beyond their designation as the English-only "control group."

Bilingual Education: Bilingual education differs fundamentally from English immersion in that it gives English language learners significant amounts of instruction in reading and/or other subjects in their native language. In the U.S., the overwhelming majority of bilingual programs involve Spanish, due to the greater likelihood of a critical mass of students who are Spanish-dominant and to the greater availability of Spanish materials than those for other languages. There are bilingual programs in Portuguese, Chinese, and other languages, but these are rare. In transitional bilingual programs, children are taught to read entirely in their native language through the primary grades and then transition to English reading instruction somewhere between second and fourth grade. English oracy is taught from the beginning, and subjects other than reading may be taught in English, but the hallmark of transitional bilingual education is the teaching of reading in the native language for a period of time. Such programs can be "early-exit" models, with transition to English completed in second or third grade, or "late-exit" models, in which children may remain throughout elementary school in native-language instruction to ensure their mastery of reading and content before transition (see Ramirez, Pasta, Yuen, Billings, & Ramey 1991). Alternatively, "paired bilingual" models teach children to read in both English and their native language at different time periods each day or on alternating days. Within a few years, the native language reading instruction may be discontinued, as children develop the skills to succeed in English. Willig (1985) called this model "alternative immersion," because children are alternatively immersed in native language and English instruction.

Two-way bilingual programs, also called dual language or dual immersion, provide reading instruction in the native language (usually Spanish) and in English both to ELLs and to English speakers (Calderón & Minaya-Rowe 2003; Howard, Sugarman, & Christian 2003). For the ELLs, a two-way program is like a paired bilingual model, in that they learn to read both in English and in their native language at different times each day.

A special case of bilingual education is programs designed more to preserve or show respect for a given language than to help children who are genuinely struggling with English. For example, Morgan (1971) studied a program in Louisiana for children whose parents

often spoke French at home, but generally spoke English themselves. These are not discussed in this paper.

Problems of Research on Language of Instruction

Research on the achievement effects of teaching in the child's native language in comparison to teaching in English suffers from a number of inherent problems beyond those typical of other research on educational programs. First, there are problems concerning the ages of the children involved, the length of time they have been taught in their first language, and the length of time they have been taught in English. For example, imagine that a transitional bilingual program teaches Spanish-dominant students primarily in Spanish in grades K-2 (kindergarten to second grade), and then gradually transitions them to English by fourth grade. If this program is compared to an English immersion program, at what grade level is it legitimate to assess the children in English? Clearly, a test in second grade is meaningless, as the bilingual children have not yet been taught to read in English. At the end of third grade, the bilingual students have partially transitioned, but have they had enough time to become fully proficient? Some would argue that even the end of fourth grade would be too soon to assess the children fairly in such a comparison, as the bilingual children need a reasonable time period in which to transfer their Spanish reading skills to English (see, for example, Hakuta, Butler, & Witt 2000).

A related problem has to do with pretesting. Imagine that a study of a K-4 transitional Spanish bilingual program began in third grade. What pretest would be meaningful? An English pretest would understate the skills of the transitional bilingual students, while a Spanish test would understate the skills of the English immersion students. For example, Valladolid (1991) compared gains from grades 3 to 5 for children who had been in either bilingual or immersion programs since kindergarten. These children's "pretest" scores are in fact posttests of very different treatments. Yet studies comparing transitional bilingual and immersion programs are typically too brief to have given the students in the transitional bilingual programs enough time to have fully transitioned to English. In addition, many studies begin after students have already been in bilingual or immersion treatments for several years.

The studies that do look at four- or five-year participations in bilingual or immersion programs are usually retrospective (i.e., researchers search records for children who have already been through the program). Retrospective studies also have characteristic biases, in that they begin with the children who ended up in one program or another. For example, children who are removed from a given treatment for systematic reasons, such as Spanish-dominant students removed from English immersion because of their low performance there, can greatly bias a retrospective study, making the immersion program look more effective than it was in reality.

Many inherent problems relate to selection bias. Children end up in transitional bilingual education or English immersion by many processes that could be highly consequential for the outcomes. For example, Spanish-dominant students may be assigned to Spanish or

English instruction based on parent preferences. Yet parents who would select English programs are surely different from those who would select Spanish in ways that would matter for outcomes. A parent who selects English may be more or less committed to education, may be less likely to be planning to return to a Spanish-speaking country, or may feel very differently about assimilation. Thomas & Collier (2002) reported extremely low scores for Houston students whose parents refused to have their children placed in either bilingual or English as a Second Language programs. Are those scores due to relatively positive effects of bilingual and ESL programs, or are there systematic differences between children whose parents refused bilingual or ESL programs and other children? It is impossible to say, as no pretest scores were reported.

Bilingual programs are more likely to exist in schools with very high proportions of English language learners, and this is another potential source of bias. For example, Ramirez et al. (1991) found that schools using late-exit bilingual programs had much higher proportions of ELLs than did early-exit bilingual schools, and English immersion schools had the smallest proportion of ELLs. This means that whatever the language of instruction, children in schools with very high proportions of ELLs are conversing less with native English speakers both in and out of school than might be the case in an integrated school and neighborhood that uses English for all students because its proportion of ELLs is low. Most problematically, individual children may be assigned to native language or English programs because of their perceived or assessed competence. Native language instruction is often seen as an easier, more appropriate placement for ELLs who are struggling to read in their first language, while students who are very successful readers in their first language or are felt to have greater potential are put in English-only classes. This selection problem is most vexing at the point of transition, as the most successful students in bilingual programs are transitioned earlier than the least successful children. A study of bilingual vs. immersion programs involving third or fourth graders may be seriously biased by the fact that the highest-achieving bilingual students may have already been transitioned, so the remaining students are the lowest achievers.

A source of bias not unique to studies of bilingual education but very important in this literature is the "file drawer" problem, the fact that studies showing no differences are less likely to be published or to otherwise come to light. This is a particular problem in studies with small sample sizes, which are very unlikely to be published if they show no differences. The best antidote to the "file drawer" problem is to search for dissertations and technical reports, which are more likely to present their data regardless of their findings (see Cooper 1998).

Finally, studies of bilingual education often say too little about the bilingual and immersion programs themselves or the degree or quality of implementation of these programs. Yet bilingual models can vary substantially in quality, amount of exposure to English in and out of school, teachers' language facility, time during the school day, instructional strategies unrelated to language of instruction, and so on.

Because of these inherent methodological problems, an adequate study comparing bilingual and immersion approaches would:

a) randomly assign a large number of children to be taught in English or their native language;

b) pretest them in their native language when they begin to be taught differentially, either in their native language or in English (typically kindergarten);

c) follow them long enough for the latest-transitioning children in the bilingual condition to have completed their transition to English and have been taught long enough in English to make a fair comparison; and

d) collect data throughout the experiment to document the treatments received in all conditions. Unfortunately, only a few, very small studies of this kind have ever been carried out. As a result, the studies that compare bilingual and English-only approaches must be interpreted with great caution.

Review Methods

This section focuses on research comparing immersion and bilingual reading programs applied with English language learners, with measures of English reading as the outcomes. The review uses a quantitative synthesis method called "best-evidence synthesis" (Slavin 1986). It uses the systematic inclusion criteria and effect size computations typical of meta-analyses (see Cooper 1998; Cooper & Hedges 1994), but discusses the findings of critical studies in a form more typical of narrative reviews. This strategy is particularly well-suited to the literature on reading programs for English language learners, because the studies are few in number and are substantively and methodologically diverse. In such a literature, it is particularly important to learn as much as possible from each study, not just to average quantitative outcomes and study characteristics.

Literature Search Strategy

The literature search benefited from the assistance of the federally commissioned National Literacy Panel on the Development of Literacy Among Language Minority Children and Youth, chaired by Diane August and Timothy Shanahan. The first author was initially a member of the Panel, but resigned in June 2002, to avoid a two-year delay in publication of the present article. This review, however, is independent of the panel's report, and uses different review methods and selection criteria. Research assistants searched ERIC, Psychological Abstracts, and other databases for all studies with the following descriptors: language minority students, English language learners, bilingual education, bilingual students, bilingualism, English as a second language, English immersion, dual language, and two-way bilingual education. Citations from other reviews and articles were also obtained. In particular, every effort was made to find all studies cited in previous reviews. From this set, we selected studies that met the criteria described below.

Criteria for Inclusion

The best-evidence synthesis focused on studies that met minimal standards of methodological adequacy and relevance to the purposes of the review. These were as follows.

1. The studies compared children taught reading in bilingual classes to those taught in English immersion classes, as defined above.

2. Either random assignment to conditions was used, or pretesting or other matching criteria established the degree of comparability of bilingual and immersion groups before the treatments began. If these matching variables were not identical at pretest, analyses adjusted for pretest differences or data permitting such adjustments were presented. Studies without control groups, such as pre-post comparisons or comparisons to "expected" scores or gains, were excluded. Studies with pretest differences exceeding one standard deviation were excluded. Those with pretest differences less than one standard deviation were included if the researchers carried out appropriate statistical adjustments.

   A special category of studies was rejected based on the requirement of pretest measurement before treatments began. These are studies in which the bilingual and immersion programs were already under way before pretesting or matching. For example, Danoff, Coles, McLaughlin, & Reynolds (1978), in a widely cited study, compared one-year reading gains in many schools using bilingual or immersion methods. The treatments began in kindergarten or first grade, but the pretests (and later, posttests) were administered to children in grades 2-6. Because the bilingual children were primarily taught in their native language in K-1 and the immersion children were taught in English, their pretests in second grade would surely have been affected by their treatment condition. Meyer & Feinberg (1992, p. 24) noted the same problem with reference to the grade 1-3 component of the Ramirez et al. (1991) study: "It is like watching a baseball game beginning in the fifth inning: If you are not told the score from the previous innings, nothing you see can tell you who is winning the game." Similarly, several studies tested children in upper elementary or secondary grades who had experienced bilingual or immersion programs in earlier years. These were included if premeasures were available from before the programs began, but in most cases such premeasures are not reported, so there is no way to know if the groups were equivalent beforehand (examples include Thomas & Collier 2002; Curiel, Stenning, & Cooper-Stenning 1980).

3. The subjects were English language learners in elementary or secondary schools in English-speaking countries. Studies that mixed ELLs and English monolingual students in a way that does not allow for separate analyses were excluded (e.g., Skoczylas 1972). Studies of children learning a foreign language were not included. However, Canadian studies of French immersion have been widely discussed, and are therefore discussed in a separate section.

4. The dependent variables included quantitative measures of English reading performance, such as standardized tests and informal reading inventories. If treatment-specific measures were used, they were included only if there was evidence that all groups fo-

cused equally on the same outcomes. Measures of outcomes related to reading, such as language arts, writing, and spelling, were not included.

5. The treatment duration was at least one school year. For the reasons discussed earlier, even one-year studies of transitional bilingual education are insufficient, because students taught in their native language are unlikely to have transitioned to English. Studies even shorter than this do not address the question in a meaningful way.

Studies that met an initial screen for germaneness to the topic, including all studies cited by Rossell & Baker (1996) or by Willig (1985), are listed in Appendix 1, which indicates whether or not each study was included and, if not, the main reasons for exclusion.

Limitations

It is important to note that the review methods applied in this best-evidence synthesis have some important limitations. First, in requiring measurable outcomes and control groups, the synthesis excludes case studies and qualitative studies. Many such descriptions exist, and these are valuable in suggesting programs or practices that might be effective. Description alone, however, does not indicate how much children learned in a given program, or what they would have learned had they not experienced that program. Second, it is possible that a program that has no effect on reading achievement measures might nevertheless increase children's interest in reading or reading behaviors outside of school. However, studies rarely measure such outcomes in any systematic or comparative way, so we can only speculate about them. Finally, it is important to note that many of the studies reviewed took place many years ago, and that both social and political contexts, as well as bilingual and immersion programs, have changed, so it cannot be taken for granted that outcomes described here would apply to outcomes of bilingual and immersion programs today.

Computation of Effect Sizes

If possible, effect sizes were computed for each study. These were computed as the experimental mean minus the control mean divided by a pooled standard deviation. When information was lacking, however, effect sizes were estimated using exact t's or p values or other well-established estimation methods (see Cooper 1998; Cooper & Hedges 1994; Lipsey & Wilson 2001). For studies lacking means and standard deviations that reported no significant difference between the experimental and control groups and did not indicate the direction of the effect (e.g., Cohen 1975), an estimated effect size of zero was used. Only English reading measures were used in determining effect sizes, even if other measures are mentioned in the text. No study was excluded solely on the grounds that it did not provide sufficient information for computation of an effect size.

Data Analysis

All data were entered into the beta version of the Comprehensive Meta-Analysis Program (Borenstein 2005) to estimate the effect sizes of each study, to calculate the overall mean weighted effect sizes, and to test whether the mean weighted effect size was derived from a homogeneous set (Q statistic). The weighting factor was sample size, so that effect sizes from larger samples contribute more to the mean than those from smaller samples. Each study contributed a single effect size to the overall mean weighted effect size. For studies that had more than one independent group or one independent outcome measure, effect sizes were calculated separately for each group and measure. These effect sizes were then weighted and averaged to create one effect size for the study. For longitudinal studies, the last time-point was used to estimate the overall effect of the study. For example, if a study followed a group of children from grade 1 to grade 5, the outcome measures for fifth graders were used to generate the effect sizes.

Previous Quantitative Reviews

The debate about empirical research on language of instruction for English language learners has largely pitted two researchers, Christine Rossell and Keith Baker, against several other reviewers. Rossell and Baker have carried out a series of reviews and critiques arguing that research does not support bilingual education (see Baker & de Kanter 1981, 1983; Baker 1987; Rossell 1990; Rossell & Baker 1996; Rossell & Ross 1986). The most comprehensive and recent version of their review was published in 1996. In contrast, Willig (1985) carried out a meta-analysis and concluded that research favored bilingual education, after controls were introduced for various study characteristics. Other reviewers using narrative methods have agreed with Willig, e.g., Wong-Fillmore & Valadez (1986). Baker (1987) and Rossell & Baker (1996) criticized the Willig (1985) review in detail, and Willig (1987) responded to the Baker (1987) criticisms.

In a review commissioned by the Tomas Rivera Center, Jay Greene (1997) carefully reexamined the Rossell & Baker (1996) review. While Rossell & Baker used a "vote-counting" method in which they simply counted the numbers of studies that favored bilingual, immersion, or other strategies, Greene (1997) carried out a meta-analysis in which each study produced one or more effect sizes, the proportion of a standard deviation separating bilingual and English-only programs. Greene categorized only 11 of the 72 studies cited by Rossell & Baker as methodologically adequate, and among these he calculated an effect size of +0.21 favoring bilingual over English-only approaches on English reading measures. Among five studies using random assignment, Greene calculated an effect size of +0.41 on English reading measures.

As part of this review, we attempted to obtain every study reviewed by Rossell & Baker and by Willig, as well as additional studies, and independently reviewed each one against the set of standards outlined previously. Consistent with Greene, we found that the Rossell & Baker (1996) review accepted many studies that lacked adequate methodology. Appendix 1 lists all of the reading studies cited by Rossell & Baker according to categories of methodo-

logical adequacy outlined in this article, which closely follow Greene's categorization. As is apparent from the Appendix, only a few of the studies met the most minimal of methodological standards, and most violated the inclusion criteria established by Rossell & Baker (1996) themselves. We found, however, that most of the 16 studies cited by Willig also do not meet these minimal standards. These are also noted in Appendix 1. In itself, this does not mean that the overall conclusions of either review are incorrect, but it does mean that the question of effects of language of instruction on reading achievement must be explored with a different set of studies than the ones synthesized by either Rossell & Baker or Willig. The Rossell & Baker and Willig studies can be categorized as follows (following Greene 1997):

1. **Methodologically adequate studies of elementary reading.** These are studies that compared English language learners taught to read using bilingual or English-only strategies, with random assignment or well-documented matching on pretests or other important variables. All of these studies focused on Spanish-dominant students.

2. **Methodologically adequate studies of heritage language programs.** Two studies, ones involving Choctaw in Mississippi and one involving French in Louisiana, evaluated bilingual programs with children who generally spoke English but were expected to benefit from introduction of their cultural language.

3. **Methodologically adequate studies of secondary programs.** We put two secondary school studies (Covey 1973; Kaufman 1968) in a separate category.

4. **Canadian studies of French immersion.** Several studies (e.g., Lambert & Tucker 1972; Genesee & Lambert 1983) evaluated French immersion programs in Canada. However, since they compared immersion to monolingual English instruction or to brief French-as-a-second-language classes, these are not evaluations of bilingual education.

5. **Studies in which the target language was not the societal language.** In addition to Canadian studies of French immersion in non-francophone areas (e.g., Day & Shapson 1988), Ramos, Aguilar, & Sibayan (1967) studied various strategies for teaching English in the Phillipines.

6. **Studies of outcomes other than reading.** A few studies (e.g., Lum 1971; Legarreta 1979; Pena-Hughes & Solis 1980) assessed only oral language proficiency, not reading.

7. **Studies in which "bilingual" treatments involved little use of native language reading instruction.** A few studies (e.g., Educational Operations Concepts 1991 a, b) evaluated programs that may be called bilingual but in fact make only incidental use of the native language, and do not use native language reading texts.

8. **Studies in which pretesting took place after treatments were under way.** As noted earlier, many studies (e.g., Danoff et al. 1978; Rosier & Holm 1980; Rossell 1990; Thomas & Collier 2002; Valladolid 1991) compared gains made in bilingual and immersion programs after the programs were well under way. Both Willig and Rossell & Baker included such studies, and Greene (1997) accepted some of them as "methodologically

adequate," but we would argue that they add little to understanding the effects of bilingual education.

9. **Redundant studies.** Rossell & Baker included many studies that were redundant with other studies in their review. For example, one longitudinal study (El Paso 1987, 1990, 1992) issued three reports on the same experiment, but it was counted as three separate studies. Curiel's 1979 dissertation was published in 1980, yet both reports were counted.

10. **No evidence of initial equality.** Several studies either lacked data on initial achievement, before treatments began, or presented data indicating pretest differences in excess of one standard deviation.

11. **No appropriate comparison group.** Many of the studies included by Rossell & Baker (1996) had no control group. For example, Burkheimer, Conger, Dunteman, Elliott, & Mowbray (1989) and Gersten (1985) used statistical methods to estimate where children should have been performing and then compared this estimate to their actual performance. Rossell & Baker's own standards required "a comparison group of LEP students of the same ethnicity and similar language background," yet they included many studies that did not have such comparison groups. Further, many studies included by Rossell & Baker lacked any information about the initial comparability of children who experienced bilingual or English-only instruction (e.g., Matthews 1979). This includes studies that retroactively compared secondary students who had participated in bilingual or English-only programs in elementary schools but failed to obtain measures of early academic ability or performance (e.g., Powers 1978; Curiel et al. 1980). Other studies compared obviously non-comparable groups. As an example of the latter, Rossell (1990) compared one-year gains of English language learners in Berkeley, California, who were in Spanish bilingual or English immersion programs, yet 48% of the ELLs, all in the English immersion programs, were Asian, while all students in the Spanish bilingual program (32% of the sample) were, of course, Latino. Also, Legarreta (1979) compared Spanish-dominant children in bilingual instruction to mainly English-dominant children taught in English. Finally, Carlisle & Beeman (2000) compared Spanish-dominant children taught 80% in Spanish and 20% in English to those taught 80% in English and 20% in Spanish, so there was no English-only comparison group.

12. **Brief studies.** A few studies cited by Rossell & Baker involved treatment durations less than one year. For the reasons discussed earlier, studies of bilingual education lasting only 10 weeks (Layden 1972) or four months (Balasubramonian, Seelye, & de Weffer 1973) are clearly not relevant. Also, all but one of these brief studies also failed to meet inclusion standards on other criteria as well (e.g., they lacked pretests or had outcomes other than reading).

The Present Review

This review carries out a best-evidence synthesis of studies comparing bilingual and English approaches to reading in the elementary and secondary grades that meet the inclusion criteria outlined above. These include the methodologically adequate studies cited in the

Willig (1985), Rossell & Baker (1996), and Greene (1997) reviews, as well as other studies located in an exhaustive search of the literature, as described previously. The characteristics and findings of these studies are summarized in Table 1.

Studies of Beginning Reading for Spanish-Dominant Students

The largest number of studies focused on teaching reading to Spanish-dominant students in the early elementary grades. Thirteen studies of this kind met the inclusion criteria.

Three categories of bilingual programs were distinguished. The most common among the qualifying studies were studies of paired bilingual strategies, in which students were taught to read in English and in Spanish at different times of the day, beginning in kindergarten or first grade and continuing through the end of the study. Pairing may not have begun on the first day of the school year, but if children were being taught to read in both Spanish and English during their first year of reading instruction, the program was considered a paired model. A second category involved evaluations of programs in which children were taught reading in Spanish for one year before a transition to paired bilingual instruction (English and Spanish). A third category consisted of a single study by Saldate et al. (1985), which did not describe the treatments well enough to permit categorization, although it seemed to evaluate a transitional model.

In Table 1, the elementary studies of Spanish-dominant children are listed according to these treatment categories, with the highest-quality studies listed first. That is, randomized multi-year studies are listed first, then matched multi-year studies, then matched one-year studies. The studies will be discussed in the same order.

Table 1: Language of Reading Instruction: Descriptive Information and Effect Sizes for Qualifying Studies

Study	Intervention description	Design	Duration	N	Grade	Sample characteristics	Evidence of initial equality	Posttest	Effect size	Mean Weighted Effect Size
Studies of paired bilingual education										
Plante (1976)	Paired bilingual	Random assignment	2 yrs	55	1-2, 2-3	Spanish-dominant Puerto Rican students in New Haven, CT	Well matched on Spanish oral vocabulary but C>E in English pretest	English Inter-American Series 2nd grade 3rd grade	 +0.78 +0.26	+0.50
Huzar (1973)	Paired bilingual	Random assignment	2 & 3 yrs	160	1-2, 1-3	Disadvantaged Puerto Rican students in Perth Amboy, NJ	Well matched on IQ, SES, and initial achievement	English Inter-American Series 2nd grade 3rd grade	 +0.01 +0.31	+0.16
Campeau et al. (1975), Corpus Christi	Paired bilingual	Matched control	2 yrs	171	K-1	Spanish dominant students in Corpus Christi, Texas	Matched on English and Spanish pretests	English Inter-American Series	+0.45	+0.45
Ramirez et al. (1991)	Paired bilingual	Matched control	2 yrs	153 students in 4 schools	K-1	Spanish dominant LEP students	Very well matched on SES and home backgrounds.	English CTBS	+0.53	+0.53
Campeau et al (1975), Houston	Paired bilingual	Matched control	3 yrs	206	K-2	Spanish dominant students in Houston, TX	Matched on language, SES, and academic achievement	English Inter-American Series	+0.54	+0.54

Table 1 continued

Study	Intervention description	Design	Duration	N	Grade	Sample characteristics	Evidence of initial equality	Posttest	Effect size	Mean Weighted Effect Size
J. R. Maldonado (1977)	Paired bilingual	Matched control	5 yrs	126	1-5	Spanish dominant students in six elementary school in Corpus Christi, TX	Matched on SES and number of years in schools	English (SRAAS)		
								2nd	0.00[a]	
								3rd	0.00a	0.00[a]
								4th	0.00a	
								5th	0.00[a]	
Alvarez (1975)	Paired bilingual	Matched control	2 yrs	147	2	Spanish dominant children in two schools in Austin, TX	Matched on SES and initial language proficiency	California Achievement Tests		
								English reading vocab	+0.12	-0.06
								English reading comp	-0.23	
Cohen (1975)	Paired bilingual	Matched control	2-3 yrs	90	K-1, 1-2, 1-3	Span. dominant students in Redwood City, CA	Matched on SES and initial language proficiency	English Inter-American Series		
								Cohort 1	0.00[a]	0.00[a]
								Cohort 2	0.00[a]	
Campeau et al (1975), Kingsville, TX	Paired bilingual	Matched control	1 yr	89	K	Spanish dominant students in Kingsville, TX	Matched on SES and ethnic mix	English Inter-American Series	+0.42	+0.42
Campeau et al (1975), Santa Fe	Paired bilingual	Matched control	1 yr	77	1	Hispanic students in Sante Fe, New Mexico	Similar on pretests, but E>C	English MAT	+0.03	+0.03

Table 1 continued

Study	Intervention description	Design	Duration	N	Grade	Sample characteristics	Evidence of initial equality	Posttest	Effect size	Mean Weighted Effect Size
Studies of one-year transitional bilingual education										
J. A. Maldonado (1994)	Bilingual-1-year transition	Random assignment	3 yrs	20	2-4, 3-5	Spanish dominant special education students in Houston, TX	Well matched on disability, language proficiency, & family background	English CTBS	+1.66	+1.66
Campeau et al (1975), Alice, TX	Bilingual-1-year transition	Matched control	2 yrs	125	K-1	Span. dominant students in Alice, TX	Similar on English pretests but E>C* on Spanish pretest	English Inter-American Series	+0.49	+0.49
Study of bilingual education (unspecified)										
Saldate et al (1985)	Unspecified	Matched control	3 yrs	38	1-3	Spanish dominant students in Douglas, AZ	Well matched on pretests	English tests MAT (2nd grade) WRAT (3rd grade)	 -0.28 +0.89	+0.89
Secondary studies										
Covey (1973)	Paired bilingual	Random assignment	1 yr	200	9	Spanish dominant students	Well matched on pretests	English Stanford Diagnostic Reading	+0.72	+0.72
Kaufman (1968)	Paired bilingual	Random assignment	1 & 2 yrs	139	7	Spanish dominant students in New York City	Initial CIA vocab and comprehension scores, language and non-language IQ, age, and Hoffman bilingual schedule scores were used as covariates	2-yr school Reading Total 1 yr school Reading Total	 +0.23 +0.23	+0.23

a Effect size estimated; data for exact computation were not available.

* E = experimental group, C = control group

Studies of Paired Bilingual Programs

Ten qualifying studies compared paired bilingual and English immersion programs. **Plante (1976)** randomly assigned Spanish-dominant, Puerto Rican children in a New Haven, Connecticut, elementary school to a paired bilingual model or to English-only instruction. Initially, 72 children were randomly assigned, 45 to the paired bilingual group and 27 to an English-only control group. By the end of the study, 31 children remained in the paired bilingual group and 22 remained in the control group. The children began in kindergarten or first grade.

The treatment involved a team-teaching arrangement with one native Spanish-speaking teacher and one English-speaking teacher. It was described as follows.

> "(Spanish dominant children) are taught their basic skills, i.e., reading, writing, arithmetic, social studies, and science, in Spanish. At the same time, the English-speaking Anglo teacher initiates the teaching of English, beginning with an aural-oral approach. When an English oral vocabulary is sufficiently developed in individual children, she initiates instruction in the reading and writing of English. The key premise in this instructional organization is a concept of diagnostic-prescriptive instruction with both Spanish and English resources being available." (Plante 1976, p. 40)

Analyses of pretest scores for the final sample found that the two groups were similar on measures of Spanish and English oral vocabulary. The control group was nonsignificantly higher on both measures.

Two years later, all children were given the English form of the Inter-American Test of Reading and the English Metropolitan Achievement Test. Second graders in the paired bilingual treatment scored significantly higher than control second graders (ES=+0.78). The effect size for third graders was also positive but nonsignificant (ES=+0.26). Total effects were nonsignificant, with a mean weighted effect size of +0.50. However, these differences did not control for the control group's pretest advantage, so pretest-adjusted differences would have further favored the experimental group.

On the Metropolitan Achievement Test, total reading scores favored the experimental group by 0.4 grade equivalents among second grade and 0.5 grade equivalents among third graders. No standard deviations or analyses were provided, however.

Finally, there were substantial differences in retention rates. Only one of the 31 experimental children (3%) was retained in grade, compared to 13 of the 22 control children (59%). Retention rates are determined by the teachers involved on a subjective basis, but this is nevertheless an important finding.

Not surprisingly, children in the bilingual group scored substantially better on a Spanish reading test than did the English-only control group (overall ES=+1.02).

The Plante (1976) study is small, and with only one class in each treatment, teacher effects were completely confounded with treatment effects. Yet its use of random assignment and a two-year longitudinal design with modest attrition makes this an important part of the research base on bilingual education.

Huzar (1973) carried out a randomized experiment involving 160 Spanish-dominant Puerto Rican children in Perth Amboy, New Jersey. The children were assigned on entry to first grade to one of two treatments, paired bilingual or control. The paired bilingual treatment was described as follows: "One bilingual teacher gave reading instruction to the class in Spanish for 45 minutes each day, while the monolingual teacher gave reading instruction in English for the same period of time each day (Huzar 1973, p. 34)."

In the control group, students were taught only in English for 45 minutes a day. "All teaching procedures, quality of materials, and time periods for reading instruction were the same, with the exception that the experimental classes received instruction in both Spanish and English, with corresponding textbooks."

There are two potential confounds in this study. First, it is unclear what the control group was doing while the experimental group received 45 minutes of daily Spanish reading instruction. It may be that the experimental students were receiving more total time in reading (English plus Spanish). Second, the English reading texts used in the two programs were different. The experimental classes used a phonetic program, the Miami Linguistic Readers. Control students used Scott Foresman's Open Highways series.

The 160 study subjects were assigned at random to four experimental and four control classes. Two classes of each treatment were at the second and third grade levels, respectively. Third graders had been in their respective treatments for three years and second graders had been in theirs for two years. Metropolitan Readiness Test first grade scores were collected from school records, and Lorge-Thorndike Intelligence Test scores were obtained for third grades, and showed no significant differences between treatment groups at either grade level.

The posttest was the English reading test of the Inter-American Series. For second graders, there were no differences (ES= +0.01). Differences at the third grade level directionally favored the experimental group (ES=+0.31), but were statistically significant for boys (E=+0.44) but not girls (ES=-0.06).

The Huzar (1973) and Plante (1976) studies are particularly important, despite taking place more than a quarter century ago. Both are multi-year experiments that, due to use of random assignment, can rule out selection bias as an alternative explanation for the findings. Both started with children in the early elementary grades and followed them for two to three years. Both used paired bilingual reading instruction by different teachers in Spanish and English, with transition to all-English instruction by second or third grade. The use of both Spanish and English reading instruction each day more resembles the experience of Spanish-dominant students in many two-way bilingual programs (see Calderón & Minaya-

Rowe 2003) than it does transitional bilingual models, which delay English reading to second or third grade.

One of the most widely cited studies of bilingual education is a longitudinal study by **Ramirez et al. (1991)**, which compared Spanish-dominant students in English immersion schools to those in two forms of bilingual education: early exit (transition to English in grades 2-4) and late exit (transition to English in grades 5-6). Schools in several districts were followed over four years. Immersion and early-exit students in four "two-treatment" schools were well matched, but a group of "one treatment" schools that implemented only bilingual or immersion treatments were poorly matched, according to the authors. Late-exit students were lower than their comparison groups in SES and their schools had much lower proportions of native English speakers. For these reasons, no direct comparisons were made by the authors between late-exit and other schools.

The "two-treatment" comparison of early-exit transitional bilingual education and English immersion is the important contribution of the Ramirez et al. (1991) study. In the immersion program, almost all teacher speech was in English at all grade levels. In the early-exit classes, teacher speech was 31% Spanish in kindergarten, 29% in first grade, 24% in second grade, 17% in third grade, and 2% in fourth grade. The early-exit program was described as follows (Ramirez et al. 1991, p. 2):

> "In an early-exit program there is some instruction in the child's primary language, 30 to 60 minutes per day. This is usually limited to the introduction of initial reading skills. All other instruction is in English, with the child's primary language used only as a support, for clarification. However, instruction in the primary language is quickly phased out over the next two years so that by grade two, virtually all instruction is in English." (Ramirez et al. 1991, p. 2)

Although the Ramirez et al. study is invariably cited as a study of transitional bilingual education, it is in fact a study of paired bilingual education, using the definition applied in the present review. The authors noted that in the early-exit kindergartens, 35.1% of instructional time was spent on English language arts and 29.9% on Spanish language arts. In first grade the corresponding numbers were 33.5% (English) and 24.2% (Spanish). In contrast, in the English immersion classes, 63.6% of instructional time was spent on English language arts in kindergarten, and 60.2% in first grade.

Percentages of time spent on math, social studies, and science, all in English in both treatments, were nearly identical in the two treatment conditions. Overall instructional time was equivalent in the two conditions, so the immersion and early-exit classes were spending about the same amount of time on language arts, as English plus Spanish time in the early-exit program was similar to English time in the immersion classes.

The longitudinal experiment had two stages. Children were pretested in kindergarten and then posttested at the end of first grade on the English CTBS. The experimental and control students were well matched on pretests, socioeconomic status, preschool experience,

and other factors. On first-grade CTBS reading, students in the bilingual group scored significantly higher than the immersion students (ES=+0.53).

The second phase of the study followed students from spring of first grade to third grade. It is not clear why the kindergarten cohort was not simply followed, although there were serious problems with attrition over the longitudinal study. However, the second-phase study does not qualify for inclusion in this review, as it uses a pretest given long after treatments had begun.

The Ramirez et al. study was so important in its time that the National Research Council convened a panel in 1991 to review it and a study by Burkheimer et al. (1989). The panel's report (Meyer & Fienberg 1992) supported the conclusions of the Ramirez et al. comparison of the early-exit and immersion programs in grades K-1. It expressed concern about the second phase on the same criterion as that applied in the present review that the pretests for the second phase were given after the students had already been in bilingual or immersion treatments for two years.

In the mid-1970s, the American Institutes of Research (AIR) produced a series of reports on bilingual programs around the U.S. (**Campeau, Roberts, Oscar, Bowers, Austin, & Roberts 1975**). These are of some interest, with one major caveat: The AIR researchers were looking for exemplary bilingual programs. They began with 96 candidates and ultimately winnowed this list down to eight. Programs were excluded if data were unavailable, not because they failed to show positive effects of bilingual programs. Nevertheless, these sites were chosen on their reputations for excellence, and a site would clearly be less likely to submit data if the data were not supportive of bilingual education. Also, the Campeau et al. (1975) evaluations were organized as successive one-year studies, meaning that pretests after the first treatment year (K or 1) are of little value. For reasons described earlier, we included only cohorts that were pretested before treatments began. With these cautions in mind, the Campeau et al. (1975) studies are described below.

A study in **Corpus Christi**, Texas, evaluated a paired bilingual program in three schools. In kindergarten, the paired bilingual treatment involved a 2-hour language period alternating English and Spanish. "In both languages, the patterned practices stress basic sentence patterns and illustrate changes in word forms and word order." Over the course of kindergarten the proportion of Spanish instruction decreased from 90% to 50%. In first grade, the paired bilingual classes had a daily two-hour English reading and language arts period and a one-hour Spanish reading and language arts period:

> "Phonetic analysis skills are taught first in Spanish because of the highly phonetic nature of the language…English readiness skills…are taught through the Harcourt Brace Jovanovich series. Phonetic analysis skills are introduced after the child has learned them though his Spanish reading lessons…There is a heavy phonics emphasis, with comprehension skills receiving heavier emphasis as decoding skills develop." (Campeau et al. 1975, p. D-60)

Classes in the control groups were "taught monolingually without regard for the language dominance of the children" (p. D-66), but otherwise instructional strategies were similar.

Experimental and control classes were pre- and posttested each year for two years. For reasons discussed earlier, we did not include comparisons in which pretests were given after treatments were already underway. For this reason, we only focused on the children who were pretested in kindergarten in the first study year and then posttested at the end of first grade in the second study year. This cohort was well matched on the Stanford Early School Achievement Test. There were no differences at the end of kindergarten, but the paired bilingual students scored significantly higher than controls on the Inter-American English Reading Test (ES=+0.45) at the end of first grade. There were also substantial differences on the SRA Achievement Test in reading. Experimental students averaged a grade equivalent of 2.3, controls 1.8. However, no statistical comparisons were made on the SRA. The paired bilingual group also scored substantially better than the control group in Spanish reading, of course.

A study in **Houston** also reported by **Campeau et al. (1975)** followed three cohorts of students in seven paired bilingual and two English immersion schools. The paired bilingual program began with a Spanish decoding program and then Spanish Laidlaw basals, during a regular Spanish arts block. Students were then taught English reading starting with a transitional program and then continuing with the Harcourt Brace Jovanovich basal series. Each class had a teacher and an aide. Two control schools, in which Spanish-dominant students were taught only in English, were selected on the basis of similarity to seven experimental schools in language, SES, and prior achievement.

Experimental and control kindergarten students were well matched on the Inter-American English Ability Test. At the end of first grade and second grade, the paired bilingual students scored substantially higher than the control students with effect sizes of +0.69 and +0.54 respectively.

Cohen (1975) compared two schools serving many Mexican Americans in Redwood City, California. One school was using what amounts to a two-way bilingual program, in that Spanish-dominant students and English-dominant students were taught in both Spanish and English. However, from the perspective of the ELLs, the treatment was the same as a paired bilingual model. Spanish dominant students were taught Spanish reading using readers such as Pepin en Primer Grado and English reading using the phonetic Miami Linguistic Series. Spanish-dominant and English-dominant students were grouped together for content area instruction but not for reading. Three successive cohorts were compared at the two schools: grades K-1, 1-2, and 1-3. In each case, students were pretested and posttested on a broad range of English reading measures. In all cohorts, Mexican-American students were well matched on English and Spanish pretests. At posttest, there were no significant differences, adjusting for pretests. The data did not allow for computation of effect sizes, so zeros were entered in Table 1.

J. R. Maldonado (1977) conducted a five-year longitudinal study on a group of Mexican American children in six elementary schools in Corpus Christi, Texas. The main purpose of this study was to examine how well the bilingual students were able to succeed in the regular education program of the school district after they had departed the bilingual program. The experimental group consisted of 47 children who had participated in the bilingual program for four consecutive years, from first to fourth grades. The control group was comprised of 79 children who had been in a regular English-only program for the same four years and grades. The study followed the two groups until they reached fifth grade, one year after the experimental group left the bilingual program. Students in the experimental group enrolled in the Title VII program titled "Apprendemos En Dos Idioma" (we learn in two languages) in which they received "a minimum of two hours per day in Spanish language instruction in the areas of language arts, reading, and mathematics and social studies" (p. 103), no specific descriptive information about the control group was provided in the study.

No statistically significant differences were found at any grade after controlling for first-grade pretests. It is important to note that teachers in both the bilingual group and the control group were bilingual. However, it is not stated how much these bilingual teachers in the control group used Spanish in their classrooms to help children who were in need for bilingual explanations. As the researcher stated, "It is highly possible that the control group bilingual teachers might have used the Spanish language for clarification of some concepts. This in turn would not only assist those students in the comprehension of those concepts but at the same time lower the difference between the groups in the areas of mathematics and reading" (p. 104).

A study by **Alvarez (1975)** followed 147 Mexican-American children in two schools in Austin, Texas, from first to second grades. Students in the bilingual classrooms and the control classrooms in each school were well matched on socio-economic status and initial language proficiency. The instruction program in the bilingual classrooms was described as "a balanced combination of Spanish and English (50 percent/50 percent)" (p. 73). However, children who had very limited English proficiency in these bilingual classrooms were initially taught in Spanish. Reading and language arts were taught in both languages each day for two hours by the English-speaking and Spanish-speaking teachers. Oral language in English was taught for one-half hour daily to the Spanish-dominant children, and oral language in Spanish was taught for one-half hour to the English-dominant children.

Students in the control classrooms had a similar curriculum to that of the bilingual classrooms. The only difference was "that the subject matter was taught completely in English; using the same textbooks for the English component of the bilingual classes, but using twice as much time as the balanced combination of Spanish and English curriculum designed for Mexican students in bilingual classrooms" (75). The bilingual students scored somewhat higher on the English reading vocabulary test but the control group scored higher on the English reading comprehension test than the bilingual classes. None of these differences were statistically significant.

Two of the studies carried out by **Campeau et al. (1975)** had one-year durations. The first study was conducted in two low SES elementary schools in Kingsville, Texas. Five grades were compared at the two schools: K-4. Only results from the kindergarten groups, 48 in the bilingual group and 41 in the control group, are interpretable because the pretests for other grades were administered after the treatments began.

All teachers in kindergarten were bilingual. Instruction time was equally divided between Spanish (50%) and English (50%). Kindergarteners with very limited English were taught primarily in Spanish until their English proficiency reached a point where they could cope with bilingual instruction.

At the end of the study, students in both groups were given the same posttests. The mean gains between the two groups were compared. Students in all six kindergarten classes gained significantly more on the English version of the Inter-American Series than their counterparts in the control group, with an effect size of +0.42.

Another one-year study in Santa Fe, New Mexico, compared paired bilingual and immersion programs for Spanish-dominant students. Pre- and posttests were reported for each year, but only first grade was interpretable, as pretests for other years had already been affected by the treatments. Parents chose to place their children in bilingual or English programs, and apparently parents of higher-achieving children chose the bilingual group, as pretest scores were higher in that group. Students in the bilingual classrooms received "a bilingual presentation of all the topics of study in the normal curriculum" (p.D-16). For example, students were taught certain concepts in Spanish in the morning and were retaught the same lesson in English in the afternoon. In addition, students in the bilingual classrooms were grouped by ability during language arts and reading periods. No specific description information was provided about the control classrooms. At the end of the study, the bilingual group gained slightly more than the control group in English reading, with an effect size of +0.03.

Studies of One-Year Transitional Bilingual Education

J.A. **Maldonado (1994)** carried out a small, randomized study involving English language learners who were in special education classes in Houston. Twenty second- and third-graders with learning disabilities were randomly assigned to one of two groups. A bilingual group was taught mostly in Spanish for a year, with a 45-minute ESL period. During a second year, half of the instruction was in English, half in Spanish. In a third year, instruction was only in English. The control group was taught in English all three years. Students in both groups received the same amount of reading and language instruction and were taught by similar teachers.

Children were pretested on the CTBS (Comprehensive Tests of Basic Skills) and then posttested on the CTBS three years later. At pretest, the control group scored nonsignificantly higher than the bilingual group, but at posttest the bilingual group scored far higher. Using

the means and standard deviations presented in the article, the effect size would be +8.33, but using the given values of t and n, the effect size is +1.67, a more credible result.

One of the **Campeau et al. (1975)** studies, in Alice, Texas, compared Spanish-dominant students in bilingual and English immersion programs starting in kindergarten, for a two-year experiment. The treatment involved teaching kindergartners in Spanish. In first grade, children were transitioned to English reading, and were then taught equal amounts of time in each language. Matched control students were taught only in English. While kindergartners were comparable at pretest on English measures of general ability, bilingual students scored substantially higher on a Spanish ability test. At posttest (controlling for pretests), bilingual students scored substantially better on the Inter-American English reading test at the end of first grade, after two years of bilingual education (ES = +0.49).

Study of Unspecified Bilingual Education

In a poorly specified study of bilingual education, **Saldate, Mishra, & Medina (1985)** studied 62 children in an Arizona border town who attended immersion or bilingual schools. The bilingual treatment appeared to be a transitional model. The children were individually matched on the Peabody Picture Vocabulary Test in first grade. At the end of second grade, the bilingual students scored nonsignificantly lower on the English Metropolitan Achievement Test (MAT) (ES=-0.28) and higher on the Spanish MAT (ES=+0.44). This was to be expected, as they had not yet transitioned to English instruction. At third grade, however, the bilingual students (who had now transitioned to English-only instruction) substantially outperformed the immersion students both in English (ES=+0.89) and in Spanish (ES=+3.01). This study's small size means that its results should be interpreted cautiously, especially as the number of pairs dropped from 31 to 19 between second and third grades.

Studies of Secondary Reading

Two qualifying studies evaluated programs that introduced Spanish-language instruction to ELLs in the secondary grades. Both of these used random assignment.

Covey (1973) randomly assigned 200 low-achieving Mexican-American ninth graders to bilingual or English-only classes. The experimental intervention is not described in any detail, but it clearly involved extensive use of Spanish to supplement English in reading, English, and math. The groups' scores were nearly identical at pretest, but at posttest the bilingual students scored significantly better on the Stanford Diagnostic Reading Test (ES=+0.72).

Kaufman (1968) evaluated a program in which low-achieving Spanish-speaking seventh graders were randomly assigned to bilingual or English-only conditions in two New York junior high schools. One school participated in the program for a year and the other for

two years. In the bilingual classes, students received three or four periods of Spanish reading instruction each week, while controls were in art, music, or health education. On standardized tests of reading, there were non-significant differences favoring the bilingual classes in the two-year school and the one-year school (ES=+0.23 for both). The secondary studies point to the possibility that providing native language instruction to low-achieving ELLs in secondary school may help them with English reading. This application is worthy of additional research (also see Klingner & Vaughn 2004).

Canadian Studies of French Immersion

There are several Canadian studies (e.g., Lambert & Tucker 1972) that have played an important role in debates about bilingual education. These are studies of French immersion programs, in which English speaking children are taught entirely or primarily in French in the early elementary years. Rossell & Baker (1996) emphasized these studies as examples of "structured English immersion," the approach favored in their review. However, Willig (1985) and other reviewers have excluded them. The Canadian studies do not meet the inclusion standards of this review because the Anglophone children are learning a useful second language, not the language for which they will be held accountable in their later schooling. Although many of the studies took place in Montreal, the children lived in English-speaking neighborhoods, and attended schools in an English system. The focus of this review is on bilingual education used to help children succeed in the language in which they will be taught in the later grades, but the French immersion children in Canada are headed to English secondary schools. Further, these studies all involve voluntary programs, in which parents wanted their children to learn French, and the children in these studies were generally upper middle class, not disadvantaged.

Because French immersion programs were voluntary, children who did not thrive in them could be and were routinely returned to English-only instruction. This means that the children who complete French immersion programs in Canada are self-selected, relatively high achievers. Most importantly, the "bilingual" programs to which French immersion is compared are nothing like bilingual education in the U.S. At most, children receive 30 to 40 minutes daily of French as a second language, with far less time in French reading instruction than a U.S. student in a bilingual program would receive in English during and after transition (see Genesee & Lambert 1983). Yet in many studies, English comparison groups were not learning French at all. In the widely cited study by Lambert & Tucker (1972), Anglophones in French immersion classes were compared to Anglophones taught only in English, and to Francophones taught only in French. Ironically, studies of this kind, cited by Rossell & Baker (1996) as comparisons of immersion and bilingual education, are in fact comparisons of immersion and monolingual education. If they existed, Canadian studies of, say, Spanish speakers learning French in Francophone schools in Quebec or English in Anglophone schools in the rest of Canada would be relevant to this review, but studies of

voluntary immersion programs as a means to acquiring French as a second language are only tangentially relevant.

While the Canadian immersion studies are not directly relevant to the question of the effectiveness of bilingual programs for ELLs learning the societal language, they are nevertheless interesting in gaining a broader understanding of the role of native language in foreign language instruction. As a group, these are matched studies of high methodological quality. Quite in contrast to U.S. studies, however, the focus of the Canadian studies is on whether or not French immersion harms the English language development of native English speakers. It is taken as obvious that French all day will produce more facility in French than 30 to 40 minutes daily in second language classes.

Overall, the Canadian studies paint a consistent picture (Lambert & Tucker 1972; Lambert & Tucker 1977; Barik & Swain 1975; Barik, Swain, & Nwanunobi 1977; Genessee & Lambert 1985; Day & Shapson 1988). At least for the overwhelmingly middle-class students involved, French immersion had no negative effect on English reading achievement, and it gave students facility in a second language. The relevance to the U.S. situation is in suggesting that similar second-language immersion programs, perhaps including two-way bilingual programs for English proficient children, are not likely to harm English reading development. However, the relevance of these studies to any context in which the children of immigrants are expected to learn the language that will constitute success in their school and in the larger society is minimal.

Comparison of Paired Bilingual and Transitional Bilingual Programs

As noted earlier, many of the programs with the strongest positive effects for English language learners used a paired bilingual approach, in which children were taught reading in both English and their native language at different times each day from the beginning of their schooling. This approach contrasts with transitional bilingual education (TBE) models in which children are first taught to read primarily in their native language, and only then transitioned gradually to English-only instruction. Only one study has compared reading outcomes of these two bilingual approaches.

A longitudinal study by Gersten & Woodward (1995) initially favored paired bilingual instruction over TBE, but later found them to be equivalent. This study was carried out with Spanish-dominant ELLs in 10 El Paso elementary schools. Five schools used a program in which all subjects were taught in English, but Spanish instruction was also provided, for 90 minutes daily in first grade declining to 30 minutes a day in fourth grade. The transitional bilingual program involved mostly Spanish instruction with one hour per day for ESL instruction, with gradual transition to English completed in the fourth or fifth grade. The children were well matched demographically on entry to first grade, and scored near zero on a measure of English language proficiency. In grades 4, 5, 6, and 7, Iowa Tests of Basic Skills were compared for the two groups. On Total Reading, the paired bilingual students

scored significantly higher than the transitional bilingual students in fourth grade (ES=+0.31), but the effects diminished in fifth grade (ES=+0.18), and were very small in sixth (ES=+0.06) and seventh grades (ES=+0.08). Tests of language and vocabulary showed similar patterns. This pattern is probably due to the fact that the transitional bilingual students had not completed their transition to English in fourth and fifth grades. When they had done so, by sixth grade, their reading performance was nearly identical.

Research comparing alternative bilingual models is far from conclusive, but nothing suggests that it is harmful to children's reading performance to introduce both native language and English reading instruction at different times each day.

Discussion

The most important conclusion from research on language of instruction is that there are far too few high-quality studies of this question. Willig (1985) and Rossell & Baker (1996) agree on very little, but both of these reviews call for randomized, longitudinal evaluations to produce a satisfying answer to this critical question. Of course, many would argue that randomized evaluations are needed on most important questions of educational practice (see, for example, Mosteller & Boruch 2002; Slavin 2003), but in bilingual education, this is especially crucial due to the many inherent problems of selection bias in this field. Further, this is an area in which longitudinal, multi-year studies are virtually mandatory, to track children initially taught in their native language through their transition to English. Finally, while randomized, longitudinal studies of this topic are sorely needed, there are simply too few experimental studies of all kinds, including ones with matched experimental and control groups.

With these concerns in mind, however, research on language of instruction does yield some important lessons at least worthy of further study. Across 17 qualifying studies of all types of programs, 12 found effects favoring bilingual education and 5 found no differences. None of the studies found results favoring English immersion.

The largest group of studies focused on elementary reading instruction for Spanish-dominant students. Nine of 13 studies in this category favored bilingual approaches, and four found no differences. The median effect size for all 13 studies was +0.45. This effect size is higher than the estimate of +0.21 given by Greene (1997), but Greene did not locate the Campeau et al. (1975) studies that added several positive effect sizes. Also, many of the largest positive effect sizes were from studies with very small sample sizes. The mean sample size-weighted effect size for the 13 studies of elementary reading for Spanish-dominant students was +0.33. Using procedures described by Lipsey & Wilson (2001), this effect was found to be significantly different from zero, $Q=29.6$, $p<.05$, $df=12$. The weighted mean for the three randomized studies was +0.62, $Q=8.53$, $p<.05$.

It was surprising to find that most of the methodologically adequate studies located evaluated forms of bilingual education quite different from those commonly used in recent

years. These are paired bilingual programs, in which children are taught to read in English and in their native language at different times each day from the beginning of their time in school. Another category of programs provided just one year of native-language instruction before transition to English-only reading. Paired bilingual strategies were used in two of the randomized studies (Huzar 1973; Plante 1976), and in a study of a one-year transitional program (J. A. Maldonado 1994). These practices contrast with practices in transitional bilingual education, in which children are typically taught to read in their native language from kindergarten to grades two, three, or four, and then transitioned to reading.

There are several reasons that paired bilingual interventions may be so prevalent among the studies reviewed. First, most of the studies reviewed took place in the 1970's, when Title VII (Bilingual Education Act of the Elementary and Secondary Education Amendment (ESEA)) was new. At that time, paired bilingual models were popular. Second, for reasons discussed earlier in this review, studies of transitional bilingual education are very difficult to do, as they should begin in kindergarten and continue past the point of transition. A four-year longitudinal study would be required to follow children from kindergarten to third grade. Allowing for student mobility, such a study must start with a large sample in order to end up with sufficient numbers of students. The U. S. Department of Education has recently funded two matched and one randomized longitudinal study to evaluate transitional bilingual education, but before these only the Ramirez et al. (1991) study had the resources to carry out an investigation of this kind, and it did not follow a consistent sample from kindergarten to third grade.

It is important to note that most of the studies that did not qualify for inclusion also used paired bilingual models, not transitional bilingual models. A key exception was a series of studies by Thomas & Collier (2002) that followed children who had been in transitional programs but lacked pretest measures from before the TBE interventions began.

Because of the dearth of studies of TBE, it is not currently possible to say with confidence whether paired bilingual models are more effective than transitional models. Only one study, by Gersten & Woodward (1995), made this comparison. It found differences favoring paired bilingual strategies in Grades 4 and 5, but not in Grades 6 and 7. However, given the support for paired bilingual methods seen in this review, it is worthwhile to speculate about why paired methods might be beneficial.

Teaching a Spanish-speaking English language learner in Spanish can be expected to establish the alphabetic principle, the idea that words are composed of distinct sounds represented by letters (see National Reading Panel 2000). Early in their reading instruction, children learn to combine letters and sounds into words they know. This process is very difficult if children must form letters and sounds into words they don't know, so it may greatly facilitate phonetic development to learn the alphabetic principle in a familiar language rather than an unfamiliar one. Once a Spanish-speaking child can confidently decode Spanish text, he or she should be able to make an easy transfer to decoding any alphabetic language, such as English, by learning a modest number of new sounds for particular graph-

emes (Lindsey, Manis, & Bailey 2003). Several of the studies of paired bilingual instruction clearly described a process of teaching Spanish reading phonetically and then planfully transferring those skills to English decoding.

Rather than confusing children, as some have feared, reading instruction in a familiar language may serve as a bridge to success in English, as phonemic awareness, decoding, sound blending, and generic comprehension strategies clearly transfer among languages that use phonetic orthographies, such as Spanish, French, and English (see August 2002; August, Calderón, & Carlo 2001; August & Hakuta 1997; Durgonolglu, Nagy, & Hancin-Bhatt 1993; Fitzgerald 1995; Garcia 2000; Lee & Schallert 1997; Lindsey, Manis, & Bailey 2003).

Only two studies of secondary programs met the inclusion criteria, but both of these were very high quality randomized experiments. Covey (1973) found substantial positive effects of Spanish instruction for low-achieving ninth graders, while Kaufman (1968) found mixed, but slightly positive, effects of a similar approach with low-achieving seventh graders.

As noted previously, research on language of instruction may suffer from publication bias, the tendency for journals to publish only articles that find significant differences. However, dissertations and technical reports (e.g., Covey 1973; Huzar 1973; Plante 1976) less likely to suffer from publication bias also tended to favor bilingual programs.

Teaching reading in two languages, with appropriate adaptations of the English program for the needs of English language learners, may represent a satisfactory resolution to the acrimonious debates about bilingual education. Proponents of bilingual education want to launch English language learners with success while maintaining and valuing the language they speak at home. Opponents are concerned not so much about the use of native language, but about delaying the use of English. Paired bilingual models immerse children in both English reading and native language reading at the same time. They are essentially half of a two-way bilingual model; by encouraging English proficient students to also take Spanish reading, any school with a paired bilingual model can readily become a two-way program, offering English-only children a path to early acquisition of a valuable second language (see Calderón & Minaya-Rowe 2003; Howard et al. 2003).

Language of instruction must be seen as only one aspect, however, of instructional programming for English language learners. As many previous reviewers have concluded, quality of instruction is at least as important as language of instruction. (For reviews of effective programs and practices for ELLs, see August & Hakuta 1997; Fitzgerald 1995; Klingner & Vaughn 2004; Slavin & Cheung 2005, in press).

Clearly, there is much more we need to know about the role of native language instruction in reading. The research reviewed in this article may represent the best experimental studies currently available, but better evidence is needed. Longitudinal experiments using random assignment of students to alternative treatments are particularly needed. Both qualitative and quantitative research are needed to illuminate the conditions under which native language instruction may be beneficial for developing English reading skills, and to explain

these effects. Research systematically varying program components and research combining quantitative and qualitative methods are needed to more fully understand how various interventions affect the development of reading skills among English language learners. It is time to end the ideological debates, and to instead focus on good science, good practice, and sensible policies for children whose success in school is so important.

References

Alvarez, J. (1975). Comparison of academic aspirations and achievement in bilingual Versus Monolingual Classrooms. Ph.D. dissertation, University of Texas at Austin.

August, D., Calderón, M., & Carlo, M. (2001). The transfer of skills from Spanish to English: A study of young learners. NABE News, May/June (11-12), 42.

August, D. (2002). English as a second language instruction: Best practices to support the development of literacy for English language learners. Baltimore: Johns Hopkins University, Center for Research on the Education of Students Placed at Risk.

August, D., & Hakuta, K. (1997). Improving schooling for language-minority children: A research agenda. Washington, DC: National Research Council.

Bacon, H., Kidd, G., & Seaberg, J. (1982). The effectiveness of bilingual instruction with Cherokee Indian students. Journal of American Indian Education, 21 (2), 34-43.

Baker, K. (1987). A meta-analysis of selected studies in the effectiveness of bilingual education. Review of Educational Research, 57, 351-362.

Baker, K., & de Kanter, A. (1981). Effectiveness of bilingual education: A review of the literature. (Final draft report). Washington, DC: Office of Technical and Analytic Systems, U.S. Department of Education.

Baker, K., & de Kanter, A. (1983). An answer from research on bilingual education. American Education, 56, 157-169.

Balasubramonian, K., Seelye, H., and Elizondo de Weffer, R. (1973). Do bilingual education programs inhibit English language achievement: A report on an Illinois experiment. Paper presented at the 7th Annual Convention of Teachers of English to Speakers of Other Languages, San Juan.

Barik, H, and Swain, M. (1975). Three year evaluation of a large-scale early grade French immersion program: The Ottawa study. Language Learning. 25 (1), 1-30.

Barik, H. and Swain, M. (1978). Evaluation of a bilingual education program in Canada: The Elgin study through grade six. Commission Interuniversitaire Suisse de Linguistique Appliquee. Switzerland.

Barik, H., Swain, M. and Nwanunobi, E. A. (1977). English-French bilingual education: The Elgin study through grade five. Canadian Modern Language Review. 33, 459-475.

Bates, E. & May B. (1970). The effects of one experimental bilingual program on verbal ability and vocabulary of first grade pupils. Ph.D. dissertation, Texas Tech University.

Borenstein, Michael (2005). Comprehensive meta-analysis software (beta-version). Englewood, NJ: BioStat.

Bruck, M, Lambert, W. E., and Tucker, G. R. (1977). Cognitive Consequences of Bilingual Schooling: The St. Lambert Project Through Grade Six. Linguistics. (24), 13-33.

Burkheimer, G. J., Conger, A.J., Dunteman, G.H., Elliott, B.G., and Mowbray, K.A. (1989). Effectiveness of services for language-minority limited-English-proficient students. Washington, DC: U.S. Department of Education.

Calderón, M. (2001). Curricula and methodologies used to teach Spanish-speaking limited English proficient students to read English. In R.E. Slavin & M. Calderón (Eds.), Effective Programs for Latino Students. Mahwah, NJ: Erlbaum.

Calderón, M., & Minaya-Rowe, L. (2003). Designing and implementing two-way bilingual programs. Thousand Oaks, CA: Corwin.

Carlisle, J.F. & Beeman, F.F. (2000). The effects of language of instruction on the reading and writing achievement of first-grade Hispanic children. Scientific Studies of Reading, 4 (4), 331-353.

Campeau, P. L., Roberts, A. Oscar H., Bowers, John E., Austin, M., and Roberts, S. J. (1975). The identification and description of exemplary bilingual education programs. Palo Alto, CA: American Institutes for Research.

Carlo, M.S., August, D., McLaughlin, B., Snow, C.E., Dressler, C., Lippman, D., Lively, T., & White, C. (2004). Closing the gap: Addressing the vocabulary needs of English language learners in bilingual and mainstream classrooms. Reading Research Quarterly, 39 (2), 188-215.

Ciriza, F. (1990). Evalution report of the Preschool Project for Spanish-speaking Children 1989-90. San Diego, CA: San Diego City Schools.

Cohen, A. D. (1975). A sociolinguistic approach to bilingual education. Rowley, MA: Newbury House Press.

Cooper, H. (1998). Synthesizing research (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H.M., & Hedges, L.V. (Eds.) (1994). The handbook of research synthesis. New York: Russell Sage Foundation.

Covey, D. D. (1973). An analytical study of secondary freshmen bilingual education and its effects on academic achievement and attitudes of Mexican American students. Ph.D. dissertation, Arizona State University.

Curiel, H., Stenning, W., and Cooper-Stenning, P. (1980). Achieved ready level, self-esteem, and grades as related to length of exposure to bilingual education. Hispanic Journal of Behavioral Sciences, 2, 389-400.

Danoff, M. N., Arias, B.M., Coles, G. J., and Others. (1977a). Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program. Palo Alto, CA: American Institutes for Research.

Danoff, M.N., Coles, G.J., McLaughlin, D.H. & Reynolds, D.J. (1978). Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Programs, Vol. III: Year Two Impact Designs. Palo Alto, CA: American Institutes for Research.

Danoff, M. N., Coles, G. J., McLaughlin, D. H., and Reynolds, D. J. (1977b). Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Programs, Vol. I: Study Design and Interim Findings. Palo Alto, CA: American Institutes for Research.

Day, E. M., & Shapson, S. M. (1988). Provincial assessment of early and late French immersion programs in British Columbia, Canada. Paper presented at the April meetings of the American Educational Research Associates. New Orleans.

de Weffer, R. (1972). Effects of first language instruction in academic and psychological development of bilingual children. Ph.D. dissertation, Illinois Institute of Technology.

Doebler, L.K. & Mardis, L.J. (1980-81). Effects of a bilingual education program for Native American children. NABE Journal, 5 (2), 23-28.

Durgunoglu, A., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. Journal of Educational Psychology, 85 (3), 453-465.

Educational Operations Concepts, Inc. (1991a). An evaluation of the Title VII ESEA bilingual education program for Hmong and Cambodian students in kindergarten and first grade. St. Paul, MN: Author.

Educational Operations Concepts, Inc. (1991b). An evaluation of the Title VII ESEA bilingual education program for Hmong and Cambodian students in junior and senior high school. St. Paul, MN: Author.

El Paso Independent School District. (1987). Interim report of the five-year bilingual education pilot 1986-1987 school year. Office for Research and Evaluation. El Paso.

El Paso Independent School District. (1990). Bilingual education evaluation: The sixth year in a longitudinal study. El Paso, TX: Office for Research and Evaluation.

El Paso Independent School District. (1992). Bilingual education evaluation. El Paso, TX: Office for Research and Evaluation.

Fitzgerald, J. (1995). English as a second language instruction in the United State: A research review. Journal of Reading Behavior, 27, 115-152.

Garcia, G. (2000). Bilingual children's reading. In M.L. Kamil, P.B. Mosenthal, P.D. Pearson, & R. Barr (Eds.), Handbook of Reading Research, Vol. III. Mahway, NJ: Erlbaum, 813-834.

Genesee, F., Lambert, W. E., and Tucker, G. R. (1977). An experiment in trilingual education. Montreal, Canada: McGill University.

Genesee, F. & Lambert, W. E. (1983). Trilingual education for majority-language children. Child Development, 54, 105-114.

Gersten, R. (1985). Structured immersion for language minority students: Results of a longitudinal evaluation. Educational Evaluation and Policy Analysis, 7 (3), 187-196.

Gersten, R., & Woodward, J. (1995). A longitudinal study of transitional and immersion bilingual education programs in one district. The Elementary School Journal, 95 (3), 223-239.

Greene, J.P. (1997). A meta-analysis of the Rossell & Baker review of bilingual education research. Bilingual Research Journal, 21 (2/3), 103-122.

Grigg, W., Daane, M., Jin, Y., Campbell, J. (2003). The nations's report card: Reading 2002. Washington, DC: US Department of Education.

Hakuta, K., Butler, Y.G., & Witt, D. (2000). How long does it take English learners to attain proficiency? The University of California Linguistic Minority Research Institute, Policy Report 2000-1.

Howard, E.R., Sugarman, J., & Christian, D. (2003). Two-way immersion education: What we know and what we need to know. Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed at Risk.

Huzar, H. (1973). The effects of an English-Spanish primary grade reading program on second and third grade students. M.Ed. thesis, Rutgers University.

Kaufman, M. (1968). Will instruction in reading Spanish affect ability in reading English? Journal of Reading, 11, 521-527.

Klingner, J. K., & Vaughn, S. (2004). Specific strategies for struggling second language readers and writers. In T. L. Jetton & J. A. Dole (Eds.), Adolescent Literacy Research and Practice. New York: Guilford.

Lambert, W. E., and Tucker, G. R. (1972). Bilingual education of children: The St. Lambert Experience. Rowley, Quebec: Newbury House.

Layden, R. G. (1972). The relationship between the language of instruction and the development of self-concept, classroom climate, and achievement of Spanish speaking Puerto Rican children. Ph.D. dissertation, University of Maryland.

Lee, J., & Schallert, D. L. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL context. TESOL Quarterly, 31, 713-739.

Legarreta, D. (1979). The effects of program models on language acquisition by Spanish-speaking children. TESOL Quarterly, 13 (4), 521-534.

Lindsey, K. A., Manis, F. R., & Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. Journal of Educational Psychology, 95 (3), 482-494.

Lipsey, M. W. & Wilson, David B. (2001). Practical meta-analysis. Thousand Oaks, CA: Sage.

Lum, J. B. (1971). An effectiveness study of English as a second language (ESL) and Chinese bilingual methods. Ph.D. dissertation, University of California at Berkeley.

Maldonado, J.A. (1994). Bilingual special education: Specific learning disabilities in language and reading. Journal of Education Issues of Language Minority Students, 14, 127-147.

Maldonado, J.R. (1977). The effect of the ESEA Title VII program on the cognitive development of Mexican American students. Unpublished doctoral dissertation, University of Houston, Houston, TX.

Matthews, T. (1979). An investigation of the effects of background characteristics and special language services on the reading achievement and English fluency of bilingual students. Seattle, WA: Seattle Public Schools, Department of Planning, Research, and Evaluation.

Melendez, W. A. (1980). The effect of the language of instruction on the reading achievement of limited English speakers in secondary schools. Ph.D. dissertation. Loyola University of Chicago.

Meyer, M.M. & Fienberg, S.E. (1992). Assessing evaluation studies: The case of bilingual education strategies. Washington, DC: National Academy of Sciences.

Morgan, J.C. (1971). The effects of bilingual instruction of the English language arts achievement of first grade children. Ph.D. dissertation, Northwestern State University of Louisiana.

Mosteller, F., & Boruch, R. (2002). Evidence matters: Randomized trials in education research. Washington, DC: Brookings Institution.

National Center for Education Statistics (2004). Language minorities and their educational and labor market indicators-recent trends. Washington, DC: U.S. Department of Education.

National Reading Panel (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Rockville, MD: National Institute of Child Health and Human Development.

Olesini, J. (1971). The effect of bilingual instruction on the achievement of elementary pupils. Unpublished doctoral dissertation, East Texas State University, Commerce, TX.

Pena-Hughes, E, and Solis, J. (1980). ABC's: McAllen's Immersion system. McAllen, TX: McAllen Independent School District.

Plante, A. J. (1976). A study of effectiveness of the Connecticut "Pairing" model of bilingual/bicultural education. Hamden, CT: Connecticut Staff Development Cooperative.

Powers, S. (1978). The influence of bilingual instruction on academic achievement and self-esteem of selected Mexican American junior high school students. Ph.D. dissertation. University of Arizona.

Ramirez, J., Pasta, D.J, Yuen, S., Billings, D. K., and Ramey, D. R. (1991). Final report: Longitudinal study of structural immersion strategy, early-exit, and late-exit transitional bilingual education programs for language-minority children. San Mateo, CA: Aguirre International (Report to the U.S. Department of Education).

Ramos, M., Aguilar, J.V., and Sibayan, B.F. (1967). The determination and implementation of language policy. Quezon City, The Phillipines: Phillipine Center for Language Study: Monograph Series 2.

Reese, L., Garnier, H., Gallimore, R., & Goldenberg, C. (2000). Longitudinal analysis of the antecedents of emergent Spanish literacy and middle-school English reading achievement of Spanish-speaking students. American Educational Research Journal, 37 (3), 633-662.

Rosier, P & Holm, W. (1980). The Rock Point Experience: A longitudinal study of a Navajo school program. Washington DC: Center for Applied Linguistics.

Rossell, C. H. (1990). The effectiveness of educational alternatives for limited-English-profficient children.In Imhoff, Gary. (Ed.), Learning in Two Languages. New Brunswick, NJ: Transaction Publishers.

Rossell, C.H. & Baker, K. (1996). The educational effectiveness of bilingual education. Research in the Teaching of English, 30 (1), pp. 7-69.

Rossell, C., & Ross, J. (1986). The social science evidence on bilingual education. Journal of Law and Education, 15, 385-419.

Rothfarb, S. H., Ariza, M.J. and Urrutia, R. (1987). Evaluation of the Bilingual Curriculum Content (BCC) Project: A three-year study, final report. Miami: Office of Educational Accountability, Dade County Public Schools.

Saldate, M., Mishra, S. P., & Medina, M. (1985). Bilingual instruction and academic achievement: A longitudinal study. Journal of Instructional Psychology, 12 (1), 24-30.

Skoczylas, R. V. (1972). An evaluation of some cognitive and affective aspects of a Spanish bilingual education program. Ph.D. dissertation, University of New Mexico.

Slavin, R.E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. Educational Researcher, 15 (9), 5-11.

Slavin, R.E., & Cheung, A. (in press). Effective early reading programs for English language learners. Language Policy and Early Literacy Education. Greenwich, CT: Information Age Publishing.

Slavin, R.E., & Cheung, A. (2005). Effective reading programs for English language learners. Baltimore, MD: Johns Hopkins University, Center for Data-Driven Reform in Education.

Stern, C. (1975). Final report of the Compton Unified School District's Title VII bilingual-bicultural project: September 1969 through June 1975. Compton City, CA: Compton City Schools.

Teschner, R.V. (1990). Adequate motivation and bilingual education. Southwest Journal of Instruction, 9 (2), 1-42.

Thomas, W.P., & Collier, V.P. (2002). A national study of school effectiveness for language minority students' long-term academic achievement. Santa Cruz, CA: University of California at Santa Cruz, Center for Research on Education, Diversity, and Excellence.

Valladolid, L. A. (1991). The effects of bilingual education of students' academic achievement as they progress through a bilingual program. Ph.D. dissertation, United States International University.

Webb, J.A., Clerc, R.J., & Gavito, A. (1987). Comparison of bilingual and immersion programs using structural modeling. Houston, TX: Houston Independent School District. Willig, A. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. Review of Educational Research, 55 (3), 269-317.

Willig, A. (1987). Examining bilingual education research through meta-analysis and narrative review: A response to Baker. Review of Educational Research, 57 (3), 269-317.

Wong-Fillmore, L., & Valadez, C. (1986). Teaching bilingual learners. In M.C. Wittrock (Ed.), Handbook of research on teaching (3rd Ed.). New York: Macmillan.

Yap, K. O., Enoki, D. Y., and Ishitani, P. (1988, April). LEP student achievement: Some pertinent variables and policy implications. Paper presented at the annual meetings of the American Educational Research Association. New Orleans.

Yeung, A.E., Marsh, H.W., & Suliman, R. (2000). Can two tongues live in harmony?: Analysis of the National Education Longitudinal Study of 1988 (NELS88) longitudinal data on the maintenance of home language. American Educational Research Journal, 37 (4), 1001-1026.

Appendix I

Disposition of Studies Reviewed

Cited by*	Authors	Remarks
Methodologically adequate--elementary reading		
RB	Alvarez (1975)	
RB	Bacon et al (1982)	
RB	Campeau et al (1975)	5 separate studies met criteria
RB & W	Cohen (1975)	
	Doebler & Mardis (1980)	
RB & W	Huzar (1973)	
	J. A. Maldonado (1994)	
RB	J.R. Maldonado (1977)	
RB	Morgan (1971)	
RB	Plante (1976)	
RB	Ramirez et al (1991)	
	Saldate et al (1985)	
Methodologically adequate--heritage languages		
	Doebler & Mardis (1980)	Choctaw in Mississippi
RB	Morgan (1971)	French in Louisiana
Methodologically adequate--secondary reading		
RB & W	Covey (1973)	
RB & W	Kaufman (1968)	
Canadian studies of French immersion		
RB	Barik & Swain (1975)	
RB	Barik et al (1977)	
RB	Bruck et al (1977)	
RB	Day & Shapson (1988)	
RB	Genesee & Lambert (1983)	
RB	Genesse et al (1989)	
RB & W	Lambert & Tucker (1972)	
Students were not learning the societal language		
RB	Ramos et al (1967)	Learning English in the Phillipines
No reading outcomes (oral language only)		
RB & W	Lum (1971)	
RB	Bates & May (1970)	6 months; no pretest data provided
RB	Elizondo de Weffer (1972)	4 months; no reading outcomes; also preference for English language usage C > E
RB & W	Legarreta (1979)	
RB & W	Pena-Hughes & Solis (1980)	Kindergarten LAS was outcome measure
RB	Ciriza (1990)	Preschool only
RB	Rothfarb et al (1987)	
Minimal use of native language instruction		
RB	Becker et al. (1982)	Not an evaluation of bilingual programs

RB	Educational Operations Concepts (1991a)	No indication that K-1 Cambodian, Hmong students were taught in native languages
RB	Educational Operations Concepts (1991b)	No indication that 7-9 Cambodian, Hmong students were taught in native languages
RB	Gersten (1985)	Study of Direct Instruction; no bilingual comparison group
Pretests were given after treatments were under way		
RB & W	Danoff, Arias & Coles (1977a)	
RB	Melendez (1980)	
RB	Olesini (1971)	
	Rosier & Holm (1980)	
RB	Rossell (1990)	
RB & W	Skoczylas (1972)	Large pretest differences.; no separate analysis for Spanish dominant students; more English dominant children in the control group
RB & W	Stern (1975)	
	Thomas & Collier (2002)	Separate studies in Maine & Houston
RB	Valladolid (1991)	
RB	Yap, Enoki, & Ishitani (1988)	
Redundant		
RB	Ariza (1988)	Redundant with Rothfarb (1987)
RB	Barik & Swain (1978)	Redundant with Barik et al (1977)
RB	Cohen et al (1976)	Redundant with Cohen (1975)
RB	Curiel et al (1980)	Redundant with Curiel (1979)
RB & W	Danoff et al (1977b & 1978)	Redundant with Danoff (1977a)
RB	El Paso ISD (1987 & 1990)	Redundant with El Paso ISD (1992)
RB	Genesee, Lambert and Tucker (1977)	Redundant with Genesee et al (1983)
RB	McConnell (1980a)	Redundant with McConnell (1980b)
No evidence of initial equality		
RB	Ames & Bicks (1978)	Large pretest difference; mixed grades and mixed languages
RB	Barclay (1969)	Large pretest differences; 7 months
RB	Bacon et al. (1982)	No measure of early academic ability
RB & W	Carsrud & Curtis (1979 & 1980)	Mixed Spanish and English dominant children in the analysis
RB	Cottrell (1971)	Poorly matched on SES. ANCOVA was used but no pretest data provided
RB	Curiel (1979)	No measure of early academic ability
RB	El Paso ISD (1992)	No measure of early academic ability
RB	Layden (1972)	Lg. pretest difference in both Spanish and English; 10 wks.
RB	Malherbe (1946)	Lacked information about initial comparability
RB	Matthews (1979)	Lacked information about initial comparability
RB	Powers (1978)	No measure of early academic ability
RB & W	Stebbins et al (1977)	No measure of early academic ability
RB	Teschner (1990)	No measure of early academic ability
RB	Vasquez (1990)	No measure of early academic ability

RB & W	Zirkel (1972)	Large pretest differences in Hartford and Bridgeport. No bilingual instruction in New Britain, New London.
No appropriate comparison group		
RB	Burkheimer et al (1989)	Compared actual performance to expected performance, no real control group
	Carlisle & Beeman (2000)	Both groups were bilingual (80-20 vs. 20-80)
RB	de la Garza & Marcella (1985)	Compared Spanish dominant to English dominant; no pretest data
RB	Lampman (1973)	Mixed Spanish and English dominant children in the pretest analysis; only separate analysis for mean gains
RB	McConnell (1980b)	Compared to a baseline group; no measure of initial comparability
RB	Medina & Escamilla (1992)	Compared Vietnamese TBE to Hispanic Maintenance Bilingual; no reading outcomes
RB	Moore & Parr (1978)	Mixed Spanish and English dominant children; also late pretests for grade 1 and 2
RB	Prewitt-Diaz (1979)	17 weeks; large pretest difference
	Thomas & Collier (1997)	No control groups
	Thomas & Collier (2002)	Separate studies in Oregon and Florida lacked control groups
	Webb, Clert, & Gavito (1987)	Scores confounded with retention status.
Brief studies		
RB	Balasubramonian et al (1973)	4 months
Unavailable		
RB & W	McSpadden (1979, 1980)	

** RB=Rossell & Baker, 1996*

W=Willig, 1985

Meta-Murky: A Rebuttal to Recent Meta-Analyses of Bilingual Education[1]

Christine H. Rossell and Julia Kuder

Bilingual education, learning to read and write in the native tongue and learning subject matter in the native tongue, is one of the most controversial educational programs in existence in the U.S., perhaps because it flies in the face of what most people think of as common sense, and because it seems contradictory to the American assimilationist imperative. Nevertheless, in the U.S., all but a handful of bilingual education programs are assimilationist and have as their goal the highest level of English language achievement that a child can achieve. In other words, bilingual education in the U.S. is different from that in much of the rest of the world in that the native tongue is typically a bridge to English not an end in and of itself. Thus, what most Americans think of as an insane idea is really not that insane as it is practiced here.

Although my reading of the tea leaves is that bilingual education is the least effective approach to educating immigrant children, the differences are not so large that an intelligent and honest person could not believe in it as the best approach to educating second language learners if they wanted to. Moreover, because it is true that it is easier to learn to read and write in your native tongue (if the native tongue is a phonetic language), there are some common sense reasons why an intelligent person would support bilingual education.

The first author has conducted several reviews of the literature to determine whether bilingual education was effective and if not, what was. The first was a limited, unsystematic review conducted in the late 1970s for the American Educational Research Association annual Review of Research in Education (Rossell 1980) for the purpose of assessing the quality of social science research introduced into educational equity court cases. Rossell (1980) concluded the research cited in court testimony in support of bilingual education was low in quantity and quality and did not demonstrate what it asserted it demonstrated. Nevertheless, the efficacy of bilingual education was still an open question.

Baker and deKanter (1981, 1983) conducted the first systematic review of the research for the Carter administration which was being sued by a school district that had been required by the federal government to create a written language for a native American tribe in Alaska so that they could be taught to read and write in their native tongue. The Carter

[1] Much of the initial work of attempting to replicate Greene's results and to determine what formulas are used when important information is missing was conducted by Bonnie Lam, a Boston University graduate in statistics and mathematics. Arun Thomas, a graduate student in mathematics, also assisted in later work.

administration's position was that bilingual education had to be provided even if the group in question did not have a written language. Someone noticed, however, that there had never been a regulatory review of the federal law funding bilingual education. Keith Baker and Adriana deKanter, social scientists working in the Department of Education, were given the task of summarizing the research as a first step in the regulatory review process. Baker and deKanter conducted an exhaustive search of the literature and concluded in 1981 that there was no evidence for the superiority of bilingual education in English language reading and math achievement compared to English language approaches to educating English language learners (ELLs). These English language approaches were sink or swim (mainstream classroom), ESL pullout (small group instruction in a pullout setting), and structured immersion (instruction in the second language in a self-contained classroom of second language learners taught at a pace the child can understand). Therefore, there was no empirical basis for the federal funding requirement for native tongue instruction.

Their comprehensive review, the largest and most systematic that had been conducted up to that point, utilized the vote count method as well as considerable narration on the quality of each study. The vote count method generally has the following steps: 1) decide which studies are scientific or reliable, 2) determine what the findings are for each scientific study both in terms of direction and statistical significance, and 3) summarize the percentage of studies finding a positive significant effect, no significant effect, or a negative significant effect for the treatments and outcomes of interest. A valid criticism of the vote count method is that each study is weighted equally. Of course, all but a handful of the reviews of the research published in refereed journals are narrative reviews that are not even as systematic as a vote count. In a narrative review, the writer has total control over which studies to summarize and the value attached to any particular study is quite idiosyncratic.

The Baker and deKanter (1981, 1983) review went against the politically correct position of that time and it was inevitable that there would be critics. One of these was Willig (1985) who conducted a meta-analysis of a sub-sample of the Baker and deKanter studies. Meta-analysis seems deceptively simple to the unitiated. An effect size is calculated from the mean outcome of the treatment group minus the mean outcome of the control group divided by some standard deviation. For all but the simplest studies with complete information, however, this turns out to be fiendishly difficult. There are many different formulas for calculating effect sizes when all the information is available and even more when all the information needed for the meta-analysis is not found in the study. In addition, in the large studies with many outcomes trying to ascertain which of these outcomes one is supposed to use to conduct an effect size is not easy. We are of the opinion that many meta-analyses drop studies because it is just not possible to compute an effect size from them, either because there is too much data and it is not clear which of the many tables and numbers should be used or there is too little data, for example, missing standard deviations, F ratios, or p values.

Thus an advantage of the vote count method is that one can often determine the outcome of a scientific study even when it is impossible to construct an effect size. Although meta-

analysis is all the rage now and anyone with enough information and energy to conduct a vote count review would probably just do a meta-analysis today, we are not sure that it is always better than the vote count method. Moreover, in complex studies with many outcomes and/or studies with insufficient data, it may be worse than the vote count method because mathematical errors in the original study and those produced in selecting certain outcomes and not others are incorporated into the effect sizes and given an importance they would not have in a vote count. Although this problem can be found in all statistical analyses, it may be worse with meta-analysis.

Analyses done with original data using multiple regression or other statistical procedures can be easily replicated from the data itself because there are statistical packages, such as SAS, SPSS and Stata, that are widely available and easily obtained that enable one to exactly replicate an analysis from the same data set. Meta-analysis, however, cannot be conducted by any of the major statistical packages and the raw data used in the studies one is analyzing is not available to the person doing the meta-analysis.

The published and unpublished meta-analyses themselves rarely give enough information for another researcher to replicate the numbers that appear in the meta-analysis. Virtually all provide only the most basic information that would be of interest to the general reader and there is little there for those who might want to replicate the meta-analysis. Finally, there is little consensus or agreement on what criteria to use in assessing which studies to include in a review (with some people arguing that all studies, scientific and unscientific be included). In short, meta-analysis is a promising and important form of research review, but it is no panacea. It is subject to the same selection biases as narrative and vote count reviews and it has additional problems.

The first meta-analysis on bilingual education was Willig's 1985 meta-analysis rebutting Baker and deKanter. This meta-analysis included only 15 of the 39 studies in Baker and deKanter 1981, but added one study (Olesini 1971) that Baker and deKanter had rejected because of the use of grade equivalents. Willig concluded that bilingual education was superior to other approaches, although Baker (1987) in turn critiqued her study and concluded that her different findings were a function of the different studies she analyzed. In particular, she excluded all the Canadian immersion programs, a common practice among supporters of bilingual education since the fact that structured immersion is always superior to programs that include the native tongue (in the Canadian programs the native tongue is generally English) is not a finding they like. The justification for excluding these studies are, of course, made on other grounds, some with merit (depending on the study), but others without. The Canadian studies are of high quality and the impressive volume of consistent results analyzing every variation in structured immersion and bilingual education that one could think of gives us confidence in the overall findings even if some individual studies must be rejected because of a lack of information or because there are no comparisons that are relevant for the issue of bilingual education in the U.S.

The next large-scale, systematic review, Rossell and Ross (1986), was funded by the Denver School District, which wanted to know whether Hispanic students should be educated separately in their native tongue. We reviewed all the different approaches to educating English language learners and, using the vote count method, concluded that there was no evidence for the superiority of bilingual education over any other technique.

In the early 1990s, Keith Baker and Christine Rossell began another systematic review of the literature. The strategy of Rossell and Baker was to begin with the studies reviewed in Baker and deKanter (1983) and Rossell and Ross (1986) and to add to them. The total number of studies and books read as of 1993 numbered above 500 of which 300 were program evaluations, in the sense that their purpose was to evaluate the effectiveness of TBE or some other second language acquisition technique. This is a fugitive literature, most of it unpublished and some of it available only by writing directly to school districts, and it consists in large part of local evaluations that do not even come close to meeting scientific standards. Unfortunately, the fact that an article is published in a peer reviewed, academic journal does not guarantee it is scientific either. Approximately 11 percent of the methodologically unacceptable studies were published in peer reviewed, academic journals.

Since the Rossell and Baker review published in 1986, there have been two meta-analyses claiming to find the opposite of that review. The first is Greene (1998) and the second is Slavin and Cheung (2004). Greene (1998) also looked at Spanish reading achievement as an outcome. We, however, are not interested in that outcome. It is indisputable and uncontroversial that a Spanish speaking child taught to read and write in Spanish will do better in Spanish reading and writing than will a Spanish speaking child taught to read and write in English. What is controversial is the notion that a Spanish speaking child taught to read and write in Spanish will do better in English than one taught to read and write in English and so that is the only outcome we examine or have ever examined.

The reviews criticizing Baker and deKanter (1981) and Rossell and Baker (1996a, 1996b) are not systematic surveys of all the literature on second language learning programs. Willig began with Baker and deKanter's sample and analyzed those that met her criteria or those that she thought had sufficient data for a meta-analysis. Greene (1998) began with the Rossell and Baker sample and rejected all but 11 of them.

Although Slavin and Cheung (2004) assert their research assistants searched all available databases for studies of second language learning, this was not an exhaustive search since they identified only four new studies since the 1996 Rossell and Baker review. In addition, the first author is in possession of all but two of the seven studies Slavin and Cheung assert are not available and no one ever contacted us to see if we had them. In short, Slavin and Cheung (2004) appear to have started with the Rossell and Baker studies and added a few additional studies of second language learning programs that their research assistants came across in researching the issue of reading for at risk elementary children.

The Rossell and Baker Methodological Approach

Each of the 300 program evaluations,[2] Rossell and Baker were able to find was assessed to determine if it addressed the relevant questions with a methodologically sound research design. Methodologically acceptable studies generally had the following characteristics:

1. they were true experiments in which students were randomly assigned to treatment and control groups;
2. they had non-random assignment that either matched students in the treatment and comparison groups on factors that influence achievement or statistically controlled for them;
3. they included a comparison group of LEP students of the same ethnicity and similar language background or a statistical control for ethnicity and language background;
4. outcome measures were in English using NCEs, raw scores, scale scores, percentiles, etc., but not grade equivalents;
5. additional educational treatments were either nonexistent or controlled for.

Analysis of covariance was by far the most common statistical method used to control for preexisting differences in nonexperimental studies. Many statisticians have serious reservations about whether this method succeeds in properly adjusting preexisting differences. Similarly there are doubts that matching students on important characteristics that influence achievement is entirely successful. Nevertheless, as do most statisticians, Rossell and Baker generally accepted these methods unless there were serious defects in their application. Rossell and Baker also accepted multiple regression where the differences between the treatment and control group were statistically controlled for. Although the treatment and control group might not be similar initially in these studies, they become similar by the inclusion of variables that put the groups on a level playing field. Again, virtually all statisticians accept this approach, although they may have reservations about how well it puts the groups on a level playing field. Indeed, multiple regression is the workhorse of the social science research and it would have been unthinkable for us to exclude studies where groups were not comparable, if the regression equation included variables that controlled for those differences. Rossell and Baker did not specifically include that in the formal list of criterion, but it was implicit in their discussion as well as the studies included. It is now added to

[2] The initial list of studies on bilingual education was obtained from a search of the Educational Research Information Clearinghouse (ERIC) documents, the Boston University, MIT, Boston College, and the Boston Public Library card catalogues, Language and Language Behavior Abstracts, and the bibliographies of other reviews of the literature. The studies actually reviewed were those that could be obtained from 1) ERIC; 2) University Microfilms International; 3) the journal and book holdings of Boston University, MIT, Boston College, and the Boston Public Library; 4) the National Clearinghouse on Bilingual Education; 5) the Center for Applied Linguistics; 6) the Department of Education, 7) the authors themselves; 8) inter-library loan; and 9) program evaluations for 1991-93 obtained by writing to school districts in the U.S. This is a fugitive literature, and not all studies are documented, nor could all documented studies be obtained.

point 3 in the list above in underlined italics to indicate this criterion was used even if it was not explicitly specified.

Since the Rossell and Baker review, we would add three additional criterion for this particular policy area that Rossell and Baker did not have at the time: 1) the studies have to be at least one school year in duration (both Greene and Slavin and Cheung have this new criterion and we agree with them); 2) if a U.S. program was called bilingual education, it has to be for Spanish-speakers (a criterion that neither Greene nor Slavin and Cheung have); and 3) the U.S. studies should be of elementary students (another criterion that Greene and Slavin and Cheung do not have) if bilingual education is one of the treatments. The additional criteria that we have added since Rossell and Baker (1996a, 1996b) are necessary because classroom observations and teacher interviews conducted by the first author indicate that in the U.S. only Spanish speakers get true bilingual education – that is, learning to read and write in their native tongue and getting subject matter in their native tongue. These classroom observations and interviews also reveal that there is very little bilingual education at the secondary level in the U.S., and what is called bilingual education rarely includes any native tongue instruction at all. In addition, since almost all secondary students already know how to read in their native tongue, the purpose of bilingual education at the secondary level is quite different from that at the elementary level. It is generally a drop-out prevention program rather than a way to achieve the highest level of English language competency.

Rossell and Baker were interested in all programs for second language learners that were scientific and they did not restrict their review just to bilingual education (as Greene, for example, did) nor did we restrict ourselves only to studies that examined reading (as Slavin and Cheung, for example, did). Table 1a shows the findings of Rossell and Baker (1996a) using the original criteria comparing transitional bilingual education to 1) "submersion," i.e. doing nothing, 2) ESL, 3) structured immersion, and 4) maintenance bilingual education--on second language (usually English) reading, language, and mathematics as demonstrated by 70 methodologically acceptable[3] studies using the original five criteria. Table 1a also shows the effect of structured immersion compared to ESL pullout. All of the studies in Table 1a are listed in Appendix 1a[4] in abbreviated citation form in the same categories as in Table 1a. They are also listed in alphabetical order in complete citation form in Appendix 2.

[3] There were two errors in the original Rossell and Baker "acceptable studies" bibliographic count which stated the N was 72 when in fact it was 70. These bibliographic errors do not affect their findings. McConnell was counted twice in the bibliographic count, but only once in the acceptable studies results table since it is the same exact same study. De la Garza and Medina was listed in the acceptable studies bibliography, but was supposed to be in the rejected studies bibliography. The study was not in the acceptable studies results table.

[4] All appendices (1a, 1b, 2, 3, 4, 5, 6, 7) are published separately online:
http://www.wz-berlin.de/zkd/aki/files/appendices_rossell-kuder.pdf

Table 1a: % of Methodologically Acceptable Studies Demonstrating Program Superiority, Equality, or Inferiority by Achievement Test Outcome*

	READING**	LANGUAGE	MATH
TBE v. Submersion (Mainstream)			
TBE Better	22%	7%	9%
No Difference	45%	29%	56%
TBE Worse	33%	64%	35%
Total N	60	14	34
TBE v. ESL Pullout			
TBE Better	0%	0%	25%
No Difference	71%	67%	50%
TBE Worse	29%	33%	25%
Total N	7	3	4
TBE v. Mainstream/ESL			
TBE Better	19%	6%	11%
No Difference	48%	35%	55%
TBE Worse	33%	59%	34%
Total N	67	17	38
TBE v. Structured Immersion			
TBE Better	0%	0%	0%
No Difference	17%	100%	63%
TBE Worse	83%	0%	38%
Total N	12	1	8
Structured Immersion v. ESL			
Immersion Better	100%	0%	0%
No Difference	0%	0%	0%
Total N	4	0	1
TBE v. Maint. BE			
TBE Better	100%	0%	0%
Total N	1	0	0

* Studies are listed in more than one category if there were different effects for different grades or cohorts.

** Oral English achievement for preschool programs.

SOURCE: C. Rossell and K. Baker, "The Educational Effectiveness of Bilingual Education," Research in theTeaching of English, 30 (1), February 1996: 1-74.

Studies are repeated in more than one category of outcome if they had different outcomes at different grade levels or for different cohorts.[5] Those not in the table are excluded because they did not assess alternative second language learning programs or they did not meet the five original methodological criteria shown above.

The percentages in Table 1a indicate the percentage of studies showing a program to be better than the alternative it is compared to, the percentage showing no difference, and the percentage showing the program to be worse than the alternative it is compared to. This is repeated for each achievement outcome--reading, language, and math. The total number of studies assessing the particular achievement outcome for each category of comparisons are shown below the percentages.

Looking at the original sample of studies, the rank order in terms of effectiveness would be structured immersion, mainstream classroom with ESL pullout, mainstream classroom with no special help, and bilingual education. However, there is no evidence to suggest that bilingual education is a disaster and this analysis shows that transitional bilingual education is better in reading than doing nothing (that is, a mainstream classroom) 22 percent of the time and no different 45 percent of the time. Thus, if a review of the literature pulled out the right sub-sample of studies from our review, it could easily conclude that bilingual education was superior to a mainstream classroom and we might have to agree that for that sample it is.

The recent meta-analyses of bilingual education research conducted by Greene (1998) and Slavin and Cheung (2004), which claim to have refuted Rossell and Baker, both added additional conditions. Not only do we disagree with most of the additional conditions they have added and the standards that they used to decide which studies were to be analyzed, but the two meta-analyses do not agree with each other on criteria or effect sizes.

The Criteria for Inclusion

Greene (1997) summarized the Rossell and Baker criteria as follows: "Studies that were determined to be methodologically acceptable had to: (a) compare students in a bilingual program to a control group of similar students; (b) statistically control for differences between the treatment and control groups or assignment to treatment and control groups had be to done at random; (c) base results on standardized test scores in English; and (d) determine differences between the scores of treatment and control groups by applying appropriate statistical tests." He omitted our argument that grade equivalent scores should not be used, but otherwise this seems a fair summary. He then added to our criteria several more criteria, some of which we now agree with or used at the time without enunciating,

[5] A cohort is a group of students that are followed across grades in their progression through school. Thus, a group of students who started kindergarten in 1960 and graduated from high school in 1974 would be one cohort. A second cohort might be a group of students who started kindergarten in 1961 and graduated from high school in 1975.

but most of which we do not agree with. Greene argued that the bilingual programs studied had to use the native tongue at least some of the time. We agree that to call a program "bilingual" should mean that it uses the native tongue at least some of the time. However, he is simply wrong when he concludes that only 11 of the 72 Rossell and Baker studies meet their own criteria. In fact, only 11 of the 72 studies meet their criteria *plus* his criteria. The same criticism can be levelled at Slavin and Cheung (2004). They introduced new criteria, most of which Rossell and Baker would not agree with, and then claimed that Rossell and Baker did not follow their own criteria.

Only Bilingual Education?

Greene excluded three studies (Becker and Gersten 1982; Campeau et al 1975; and Webb, Clerc, and Gavito 1987) on the grounds that the students were not in bilingual education. Rossell and Baker, however, were interested in the whole panoply of second language learning programs and so they also compared structured immersion to ESL and compared transitional bilingual education (also called early exit bilingual education) to maintenance bilingual education (also called late exit bilingual education).

Greene rejected Campeau, et al, (1975) as a study of bilingual education, but we disagree and so do Slavin and Cheung. Campeau et al. is clearly a study of bilingual education programs across the U.S. as noted in its title, "The Identification and Description of Exemplary Bilingual Education Programs." Campeau et al. found bilingual education to be better than a mainstream classroom as was noted in the Rossell and Baker (1996a) review, although only the Corpus Christi study was accepted as scientific. Greene's dismissal of Webb, Clerc, and Gavito (1987) as a study that is not of bilingual education is equally inexplicable. The title of that paper is "Comparison of Bilingual and Immersion Programs." The study is of Spanish speakers in Houston. Slavin and Cheung just ignore the study – neither including it nor specifically excluding it.

English Only Comparison?

Greene also claimed to exclude studies where the comparison (non-bilingual education) students were not taught completely in English. Again, we were interested in the whole panoply of second language acquisition studies, not just bilingual education compared to nothing. Requiring that the comparison group could have no native tongue instruction has the effect of eliminating the structured immersion programs since most use at least some native tongue and all of the Canadian ones do beginning around second grade. He justifies this on the grounds that the purpose of his review is to assess the potential benefit of Proposition 227, which makes the default assignment for English Learners a structured immersion classroom and he argues prohibits all native tongue instruction. Proposition 227, however, states only that the language of the structured immersion classroom is

"overwhelmingly" English. Overwhelmingly means not entirely and California school districts have interpreted this to mean up to 30 percent native tongue instruction is allowed.

But Greene did not consistently exclude studies where the students were not taught completely in English. He included the Ramirez study, for example, despite the fact that all the teachers in the structured immersion programs were bilingual and used at least some native tongue. Indeed, the Ramirez study notes that many of the structured immersion programs used more native tongue than the transitional bilingual education programs.

Canadian Studies of French Immersion

Almost every supporter of bilingual education wants to get rid of the Canadian studies of French immersion. They are of very high quality and the many studies assess virtually every variation one can think of in structured immersion and bilingual education. The findings are troubling to supporters of bilingual education because they show that structured immersion is always better than any second language learning program that includes the native tongue *if* one's goal is the highest level of achievement in the second language that a child is capable of.

Initially, Greene wanted to get rid of the Canadian studies because they were in foreign countries. After numerous email exchanges in which the first author argued that to eliminate foreign country studies made no sense since brains don't differ from country to country, he apparently decided to eliminate them individually on other grounds.

Slavin and Cheung (2003) eliminated the Canadian studies on the grounds that they did not have the appropriate comparison group and thus were not studies of bilingual education. Slavin and Cheung (2004) changed their reason for eliminating the Canadian studies. The new reason was that the students were not learning the dominant language of the country and the programs were interested in how well the students were doing in English. The latter problem is true of some of the French immersion studies. This is why of the dozens of Canadian immersion studies we reviewed, only six made it into our review. For a Canadian immersion study to be included in our review, it had to compare the achievement of second language learning students while they were in the French immersion program or the bilingual portion of the program and to make a comparison that could be translated into American program terms. Many of the so-called French immersion programs were in fact bilingual education, although in Canada they would call it delayed immersion or partial immersion.

Although Rossell and Baker were unable to use many of the studies of French immersion programs because they couldn't figure out how to translate them into American programs or terms or because they seemed redundant or had inadequate information or controls, the entire body of work presents consistent and clear evidence that there is a strong positive relationship between the amount of instruction in a second language and achievement in that second language.

Interestingly, although Slavin and Cheung (2003) criticize the French immersion studies in the Rossell and Baker review as not being of bilingual education, others have criticized Rossell and Baker because all of the French immersion programs became bilingual after second grade. Although this is true, it does not necessarily invalidate our use of them since we only used findings for structured immersion when the outcome for one group was from the time period when they had total French immersion (structured immersion). It is irrelevant what was going to happen to them in the future if the outcome was from the past. Others have criticized the Canadian French immersion programs because the second language learners were middle class. However, when the treatment group was middle class so was the control group. Furthermore, when the experiments were conducted with working class children, they produced the same or better results (Tucker, Lambert and d'Anglejean 1973; Bruck, Jakimak, and Tucker 1971; Cziko 1975; Genesee 1976).

Slavin and Cheung's (2004) rejection of the Canadian immersion studies because the students were not learning the dominant language of the country makes no sense at all to us. As far as we are concerned, it makes the Canadian Immersion studies stronger, not weaker, since the program outcomes are less likely to be contaminated by language being learned outside the school. In short, the Canadian studies are closer to a controlled experiment than any studies conducted in the U.S. since in the U.S. there is no way to tell how much English the children in bilingual education are getting outside the school. In addition, the Canadian researchers kept meticulous records of exactly how much of each language was being used in the programs, something that is rarely found in the American studies.

It would, however, be difficult to conduct a meta-analysis of many of the Canadian Immersion studies. Several might have to be dropped because of a lack of statistical information that could be used to construct an effect size. We have yet to attempt a meta-analysis of them, but just reviewing the studies again for this paper has given us an upset stomach. We do not look forward to trying to construct an effect size from the hundreds of outcomes reported in the six books and articles Rossell and Baker included in their review.

One Year Criterion

Greene only included studies that measured the effects of bilingual programs after at least one school year. Slavin and Cheung (2004) appear to use a similar standard. With the benefit of hindsight, we agree that the additional criterion of one school year in length is a good one. Rossell and Baker should not have accepted the authors' claim that effects would be immediate in these short-term programs. Imposing this criteria excludes five studies with different findings. They are listed in Appendix 2 with (3) after them. These studies are Barclay which found a positive effect for bilingual education in reading; Layden which found a negative effect for bilingual education in reading, but no difference in math; Balasubramonian et al. which found no difference between bilingual education and ESL; Bates which

found no difference for math, but TBE was worse in reading; and de Weffer[6] which found no difference in both reading and math. In other words, this additional criterion should have no effect on Rossell and Baker's conclusions.

Other Controls Besides Pretest

Greene only included studies that not only controlled for prior test scores, but also had an additional control for individual demographic factors that influence test scores such as family income, parental education, etc. This group of studies is labeled in Appendix 3 "Studies Excluded Because They Inadequately Control Differences Between Bilingual and English-Only Students." The requirement to have a control variable other than a pre-test is a preposterous requirement, particularly for ELL students. There is no variable more important than the pretest test score. In general, if you have a pretest, you do not need additional individual demographic controls since those variables will add little to the explained variation. Indeed, the requirement that additional demographic controls be included would eliminate most educational studies in refereed journals. Moreover, family income or parental education is not an important variable for new immigrants to a country since immigration usually means at least a temporary decline in socioeconomic status. I have asked many social scientists whether they would reject a study solely because its only control variable was the pretest score and I have found no one who would. Furthermore, Slavin and Cheung (2003; 2004) do not agree with this standard since they have numerous studies in their review that have only a pretest as a control variable.

No Appropriate Comparison Group and No Evidence of Initial Equality

As shown in Appendix 4, Slavin and Cheung (2004) reject eight of the Rossell and Baker studies for not having an "an appropriate control group."[7] However, all but one of these studies had a comparison group that was either similar or made similar by statistical analysis. The exception is the Medina and Escamilla study. It should be rejected because it compared Hispanic students to Asian students in different programs, but did not control for the ethnic difference. This is particularly a problem because we now know that the ethnic difference means the program label "bilingual" cannot be trusted.

---

[6] This author's complete name is Rafaela del Carmen Elizondo de Weffer and there is no agreement in the literature on exactly what her last name is. Dissertation abstracts shows her last name as Weffer. We believe it is de Weffer. Greene opts for de Weffer in one citation, but then changes it to Elizondo de Weffer when she is co-author of the Balasubramonian study. Slavin and Cheung have also opted for Elizondo de Weffer.

[7] Although nine studies are listed as RB, the de la Garza study is an error caused by an error in the Rossell and Baker bibliography of acceptable studies, compounded by the Slavin and Cheung failure to check the study with the results tables.

Slavin and Cheung also rejected 13 studies because there was no evidence of initial equality. However, Rossell and Baker did not have this criterion nor do most social scientists. Although it is a stronger study if the groups are initially equal, the standards of social science allow for somewhat unequal groups before the treatment if their inequality is statistically controlled for. This is not a perfect solution, but it is generally considered a reasonable one by social scientists. Indeed, the number of articles and books that would be published if this standard were applied would decline dramatically.

Slavin and Cheung's characterization of Gersten 1985 as not having an appropriate control group is incorrect and mystifying. Table II of Gersten (1985) clearly shows the experimental group (Asian students in structured immersion) and the control group (Asian students in bilingual education.) Although we now believe there is an error in the study's program labels and that the so-called bilingual education students are actually ESL pullout students,[8] there are still two appropriate comparison groups—Asians in structured immersion (a program for second language learners) versus Asians in ESL pullout (another program for second language learners). We believe this is an appropriate comparison.

Slavin and Cheung's characterization of Burkheimer et al. is equally mystifying, although Greene similarly characterizes it as not having an appropriate control group. Burkheimer et al. is a very sophisticated multiple regression analysis controlling for many instructional variables including the amount of instruction in Spanish. The only students studied were limited English proficient Spanish speakers. This is one of the highest quality and most sophisticated studies we examined. Slavin and Cheung appear to rely on Greene's evaluation and on that of Meyer and Feinberg (1992) editors of a National Academy of Science book. The latter book assesses both the Burkheimer, et al. and Ramirez, et al. studies and is critical of both. Indeed, they are only slightly more critical of the Burkheimer study than of the Ramirez study and yet both Greene and Slavin and Cheung accepted the latter. In Rossell and Baker, Burkheimer's findings appear in both the TBE worse and TBE better category as some outcomes favored bilingual and some did not. This is often cited as an advantage of meta-analysis—that is, that the effects would be averaged in a meta-analysis—but it can also be thought of as a disadvantage since it would obscure some important information.

Slavin and Cheung (2004) also allege that we relied on 14 studies that lacked any information about the initial comparability of children who experienced bilingual or English-only and they cite Matthews (1979) as the one example. The students in Matthews (1979) were matched on a great number of important variables. The problem with the Matthews study, however, is that there are no numbers in the study. It appears to have been a well designed, well thought out study, but the design and effects are described verbally so an effect size cannot be constructed from this study. That did not stop Rossell and Baker from using it in

[8] Russell Gersten now agrees that the district undoubtedly mislabeled their ESL program as a bilingual program, a fairly common occurrence for the non-Hispanic second language learning programs. Personal communication with first author 11/12/2004.

their vote count, but it would certainly stop someone from using it in a meta-analysis. Indeed, many studies were probably rejected by Greene (1998) and Slavin and Cheung (2004) because of a lack of quantitative information, but they prefer to claim something more odious about the studies. We say this because there is no category in either paper for "lack of quantitative data." Yet there are at least several studies we included in our review whose research design and findings are only described verbally and who would have to be rejected for lack of quantitative data.

There is another reason, however, why we would no longer include the Matthews study in a review, even if it had sufficient quantitative data. This study compares Asian students in bilingual education to Asian students in ESL and we no longer believe Asian students receive true bilingual education. Nor is it clear from the study exactly what treatment the Asian students are getting.

Slavin and Cheung (2004) also allege that Legaretta compared Spanish-dominant children in bilingual instruction to mainly English-dominant children taught in English. We do not understand this characterization. According to the study, 95 percent of the students in the study spoke Spanish outside the home. They also rejected Legaretta because there were no reading outcomes. This is, of course, because they were interested in the effect of bilingual education on reading, not on any other skill tested in English. Greene and Rossell and Baker, however, were interested in all outcomes.

Studies in Which the Target Language Was Not the Societal Language

Slavin and Cheung (2004) offer this criterion for rejecting studies. We see no reason to exclude these studies, although it would be another way to get rid of the Canadian French immersion studies. As noted above, we believe the effect of second language learning programs is clearer when the target language is not the societal language since what goes on in school is not confounded by what goes on outside.

Studies of Outcomes Other Than Reading

This is a criterion of Slavin and Cheung, but not of Greene nor of Rossell and Baker. We obviously had a broader goal—to evaluate all the quantitatively measured educational outcomes of second language learning programs. We had no reason to restrict ourselves only to reading.

Studies in Which Pretesting Took Place After Treatments Were Underway

Slavin and Cheung (2004) offer this criterion for rejecting studies, but we do not agree nor does Greene. Very few studies have measures of achievement before the treatment since

for programs that start in kindergarten or first grade such a measure would have to be oral or some sort of nonverbal intelligence test that is difficult to administer and that might not be comparable to the post-test. The standards of social science research only require that there be a pretest at some point and that progress after that point be tracked controlling for the pretest.

In other words, the evaluation is of change over time while in a treatment rather than just before and after a treatment.

If Slavin and Cheung had consistently applied this criterion, they would have had to limit their analyses to the following programs: 1) English reading is taught simultaneously with Spanish reading, 2) students were already proficient in English, 3) a Spanish test of achievement is the pretest and thus not comparable to the post-test, 4) a nonverbal IQ test is the pretest and thus not comparable to the post-test, or 5) the students began the bilingual education program in later grades. As shown in Appendix 4, however, most of the studies they found methodologically acceptable did in fact have pretests given after the treatment was underway. In short, they were inconsistent.

Slavin and Cheung (2003) disagree with Greene (1998) on the Rossell (1990) study. Rossell (1990) found that bilingual education was no different from ESL in the first year and inferior in the second year. Greene thinks it is an acceptable study. Slavin and Cheung (2003) argued that it did not have an appropriate comparison group because 48 percent of the English language learners were Asian. Greene (1998) and Rossell and Baker (1996a) accepted Rossell (1990), despite the fact that 48 percent of the ELLs were Asian, because Asian ethnicity was a control variable in the multiple regression equations, thus explicitly controlling for that difference.

Slavin and Cheung (2004) changed the reason for rejecting Rossell (1990). The latest reason for rejecting this study is that pretests were given after treatments were under way. Again, we think it is perfectly acceptable to measure progress over time while in the treatment and so do most social scientists, including Greene (1998). Moreover, as noted above, Slavin and Cheung inconsistently apply this standard.

Missing Studies

Greene states he could not find five studies. The first author however, has three of the missing five studies and had been providing Greene with all of the studies that he had asked for. He either neglected to ask for these or he lost them. As noted above, Slavin and Cheung's research assistants were able to find the studies Greene could not find, but failed to contact the first author of this paper for copies of the five studies they state are unavailable, but which we have (Ciriza 1990; Educational Operations Concept 1991a, 1991b; Peña-Hughes and Solis 1980; and Teschner 1990).

Redundant Studies

There are 15 studies in the Rossell and Baker review that Greene says are redundant and 10 that Slavin and Cheung (2004) say are redundant.[9] Most of the supposedly redundant studies found no difference between submersion (mainstream classroom) and TBE, but I disagree that all of these studies are redundant.

Neither Greene (1998) nor Slavin and Cheung (2003, 2004) specified *why* they thought the studies were redundant. We can only surmise that they believe a study is redundant if the study is another evaluation of the same school district even if it is different students, different schools, and different years. We disagree with this. Most of the studies in Greene and Slavin and Cheung's (2003) meta-analyses had multiple outcomes for different grades and sometimes different years. Averaging multiple outcomes for different grades within a single school or district is the same thing as averaging studies of different years or grades in the same school or district. There really is no important difference.

We believe that a study is only redundant if it is of the exact same students in the same year with the exact same tests. Using that standard, Rossell and Baker made two errors. Ariza is redundant with Rothfarb et al. (1989) and Curiel (1979) is redundant with Curiel et al., (1980) because both are of the exact same students in the same year, although the authors are different and the data is presented differently. Here is where a meta-analysis would have helped prevent these two errors since we would have obtained the same effect size and thus might have been alerted to our error.

It should be noted that Greene too made errors in counting three studies (McConnell 1980a, 1980b; Danoff et al. 1977a, 1977b; Danoff et al. 1978a, 1978b) as redundant that in fact were not counted twice. This can be seen by comparing Appendix 3 (Greene's list of Studies and Reasons for Rejection) to Table 1a, the original table from Rossell and Baker (1996a) which show the studies are only counted once.

Reanalyzing Greene's Sample

Let us assume for the moment that Greene's standards and their application to the Rossell and Baker sample are correct and that Rossell and Baker are wrong. We still cannot conclude from his sample that bilingual education is superior to a mainstream classroom or to structured immersion. For one thing, only one of the studies in the Greene sample and in the Slavin and Cheung sample includes structured immersion. That study, Ramirez, et al. found no significant difference between bilingual education and structured immersion, but it also has some biases that favor bilingual education which we discuss below in the section on testing rates.

---

[9] Slavin and Cheung (2003) assert that "It is important to note that all of these duplicate citation studies found results claimed by Rossell and Baker to favor immersion over bilingual education." This is not true as one can see by looking at Appendix 3.

In addition, Greene made an important error in summarizing his effect sizes. He did not weight the effect sizes nor the Z scores as Rosenthal (1991) and others recommend. There may be other errors. We have tried to replicate his effect sizes and Z scores and it was seldom possible to do so exactly and there were large differences in the Z scores. The formula for Hedge's g is: $\frac{(\overline{Xe} - \overline{Xc})}{Spooled}$

(the mean of the experimental group minus the mean of the control group divided by the pooled standard deviation) which seems like a simple formula except for the fact that none of the studies actually had a pooled standard deviation and most lacked a standard deviation of any kind. This is in fact why Rossell and Baker (1996a) decided not to do a meta-analysis—there was too much missing data in too many studies. Since then we have learned that this is no longer considered an obstacle and there are many "estimation" techniques that are apparently acceptable, although some seem questionable to us.

Most studies had several outcomes or means for different grades or years and a number of studies had hundreds of outcomes. Greene gives no information on the following:

- how Hedge's g is calculated from the many means that appear in each study
- how a Z score is calculated for each individual study, particularly when important information is missing
- how the pooled standard deviation is calculated
- how the pooled standard deviation is calculated when standard deviations are missing from the study
- how to compute an effect size from a multiple regression equation that has b coefficients, not adjusted means.

The reader is merely referred to Rosenthal (1991), which answers only the first question and even that not completely since Rosenthal does not give the formula for the pooled standard deviation or Z score nor does he specify what to do when important information is missing. After seven years, Greene understandably does not remember how he calculated the effect sizes, pooled standard deviations, Z scores, or what formulas he used when important data was missing other than to state he used Rosenthal. He apparently kept no notes or didn't want to take the time to look for them when contacted.

Although Greene asserts he used all the data in a study, a benefit he claims for meta-analysis, he inconsistently applied this standard. In Rossell, 1990, for example, he only used the outcomes in the year there was no significant difference, 1986-87. He ignored the outcomes in the next year when bilingual education did worse than a mainstream classroom. Similar omissions were found in a few other studies.

Table 2: A Comparison of Greene's Original Summary Table to Results When Effect Sizes and Z Scores are Weighted

		Greene's Original Table 2 with Reading Z Score Corrected			**Greene's Table 2 English Results Weighted**			**Greene's Table 2 English Results Weighted – Elementary Spanish Speakers**		
		All Tests in English	**Reading (in English)**	**Math (in English)**	**All Tests in English**	**Reading (in English)**	**Math (in English)**	**All Tests in English**	**Reading (in English)**	**Math (in English)**
Benefit of Bilingual Programs in Standard Deviations (Hedge's g)		0.18	0.21	0.12	0.03	0.00	cannot calc. from Greene	0.00	-0.06	cannot calc. from Greene
z - score		2.14	2.46	1.65	0.12	0.74	cannot calc. from Greene	-0.29	-1.28	cannot calc. from Greene
p – value <		0.05	0.05	0.1	0.45	0.23	cannot calc. from Greene	0.39	0.10	cannot calc. from Greene
95% Confidence Interval	lower	0.14	0.17	cannot calc. from Greene	-0.04	-0.07	cannot calc. from Greene	-0.09	-0.14	cannot calc. from Greene
	upper	0.22	0.24		0.11	0.08		0.08	0.03	
Significance		Statistically signif.	Statistically signif.	**Not significant**	**Not significant**	**Not significant**	cannot calc. from Greene	**Not significant**	**Not significant**	cannot calc. from Greene

Table 2 compares Greene's aggregate effect sizes and Z scores (a Z score at or above 1.96 is significant at the .05 level) from Greene's original table to the same aggregate effect sizes and Z scores weighted by sample size. We were inspired by Gersten, Baker, and Otterstedt (1998) who first pointed out that Greene had not weighted the effect sizes or Z scores. Gersten, Baker, and Otterstedt (1998) weighted Greene's individual effect sizes and computed 95% confidence intervals for English and reading (not possible for math since Greene gives us no individual study math scores) and found no significant effect for 1) elementary studies only, 2) elementary studies with random assignment, 3) all grade levels of Spanish bilingual program. In other words, all the confidence intervals included zero.

We have done some additional analyses in Table 2 that Gersten, Baker, and Otterstedt (1998) did not do. We calculated the weighted effect size,[11] the weighted Z score using the formula from Rosenthal,[12] and the 95% confidence intervals for outcomes in English (i.e. ignoring Spanish outcomes) for *all* of Greene's original sample, and for elementary Spanish bilingual education programs.[13]

The first column in Greene's table took us a long time to figure out. It is not explained in his paper. It is simply labeled "All tests in English," but it is neither an average nor a sum of all reading and language tests administered in English. After months of assuming some error had been made, we now realize that it is the average of all tests in English *including math*. We have never seen this before. The first author has been reading studies of bilingual education for about 30 years and has never seen anyone combine math, reading, oral, and language (English) scores before. It is a level of aggregation that we believe is simply inappropriate.

We have corrected a small error in the reporting of the Plante study. Greene has a positive effect size, but a negative Z score when in fact the two are supposed to agree with each other in direction. If we change the sign of the Z score for that study to a positive sign to agree with the effect size, his summary Z scores in Table 2 for reading are correct (otherwise the Z score would be 1.62).

What is amazing about Greene's report, is not just its brevity and lack of information which probably sets a new record, but the fact that individual Spanish achievement scores are reported for each study, but math scores are not. To repeat, this is amazing because no one disputes that learning in Spanish produces higher achievement in Spanish, but there is quite a bit of controversy over whether it is better to learn math in English or in the native tongue. As a result of his failure to show the math effect sizes and Z scores for individual

---

[11] The formula for the weighted mean effect size is $\sum W(ES)/\sum W$ where W is the weight and ES is the effect size. The formula for the weights is $W = (2(N_E + N_c) \times N_E N_c)/(2(N_E + N_c)^2 + N_E N_c(ES)^2)$ from Cooper, 1989 where E=the experimental group, C=the control group.

[12] The weighting of the Z scores is the sample size times the Z score, summed, and divided by the square root of the sum of the squared sample sizes (see p. 69 of Rosenthal, 1991). Some formulas use degrees of freedom instead of the sample size, which would give similar results.

[13] The formula for the confidence interval is $= \sum W(ES) / \sum(W) \pm 1.96(\sqrt{V})$ where $V = 1/ \sum W$ from Shadish and Haddock (1994): 268.

studies, we cannot weight his math effect sizes and Z scores since his individual study data is needed to do that.

The three columns on the left of Table 2 show "all tests in English" and reading to be statistically significant. However, as noted above all tests in English includes language tests, reading tests, oral tests, and math tests. Although this inappropriately aggregated outcomes is statistically significant by his standards, an effect size of .18 is not important. Nor is the reading effect size of .21 which is also statistically significant. A generally accepted rule of thumb is that .8 is a large effect, .5 is a medium effect, and .2 or smaller is a small effect (Cohen 1988, Lipsey and Wilson, 2001: 147).

The middle three columns in Table 2 show our recalculation of Greene's effect sizes accepting his data with the only correction being the sign change for the Z score for Plante. After weighting his effect sizes and Z scores, no outcomes are statistically significant.

The three columns on the far right show the weighted effect sizes for the programs where the bilingual education subjects were elementary Spanish speakers. Again, no outcomes are statistically significant.

Table 3 compares each of Greene's effect sizes to our effect sizes, also using Hedge's g. Greene's sample sizes generally do not match the sample sizes we found in these studies and so our weights are based on different sample sizes. The numbers on the left in the treatment and control columns are Greene's sample sizes and the numbers in the right are the ones we found in these studies. In some cases, there are large disparities.

Our Hedge's g effect sizes, shown in the columns labeled Rossell/Kuder used formula 1 in Appendix 7 from Table B10 in Lipsey and Wilson (2001) to calculate a pooled standard deviation when the standard deviation for each group was given in the study. If the standard deviation for each group was missing, but the standard deviation for the whole sample was included in the study, formula 14 in Appendix 7 was generally used. In some cases, such as Powers, we were only given an ANOVA table with sums of squares instead of standard deviations. From this output, we computed the pooled standard deviation as the square root of the residual mean squares. When there were outcomes for different tests, grades, or groups of experimental students in different years, the effect sizes for each group or grade were weighted and combined to create an overall effect size for the study.

For studies that used multiple regression, the numerator for the effect size is the b coefficient for the treatment group (see Equation 13 from Table B10 of Lipsey and Wilson, 2001, in Appendix 7). The effect size is $2t/\sqrt{N}$ where t is the b coefficient divided by the standard error of the b.

Table 3: A Comparison of Greene's Effect Sizes for Individual Studies to Rossell & Kuder's Effect Sizes

	Greene "All Tests in English"		**Rossell/Kuder English or Language**		**Greene Reading**		**Rossell/Kuder Reading (includes oral)**		**Rossell/Kuder Math**		**Treat-ment**	**Control**	**Std. Dev. Re-ported?**	**Random Assignment**	**Elem. Spanish**
Study	**ES**	**Z**	**ES**	**Z**	**ES**	**Z**	**ES**	**Z**	**ES**	**Z**	**N (G/RK)**	**N (G/RK)**	**Yes**	**Yes**	**Yes**
Bacon et al.,1982	0,79	2,39	No data in study		0,68	2,07	**0,70**	**3,29**	**0,91**	**4,40**	18 / **35**	18 / **18**			
Covey, 1973	0,34	2,94	**0,37**	**2,37**	0,74	4,87	**0,66**	**4,69**	**0,28**	**1,56**	86 / **90**	86 / **89**	Yes	Yes	
Huzar, 1973	0,18	0,83	No data in study		0,18	0,83	**0,16**	**1,00**	No data in study		43 / **84**	43 / **76**	Yes	Yes	YES
Powers, 1978	0,00	0,01	No data in study		-0,33	-1,53	**-0,35**	**-2,13**	**-0,06**	**-0,63**	44 / **84**	43 / **84**			
Danoff et al., 1977a	-0,03	-0,39	**-0,04**	**-1,20**	-0,12	-1,50	**-0,10**	**-2,82**	**0,12**	**3,73**	955 / **1481**	523 / **3687**			YES
Kaufman, 1968	0,20	0,72	No data in study		0,20	0,72	**0,23**	**1,10**	No data in study		43 / **51**	31 / **44**		Yes	
Plante, 1976	0,52	1,34	No data in study		0,52	1,34	**0,51**	**1,76**	No data in study		16 / **31**	12 / **22**	Yes	Yes	YES
Ramirez et al., 1991	0,01	0,08	**-0,08**	**-0,37**	0,12	0,73	**-0,15**	**-0,67**	**0,17**	**0,77**	88 / **197**	160 / **191**	Standard Error		YES
Rossell, 1990	-0,01	-0,03	**-0,24**	**-2,20**	-0,05	-0,20	**-0,25**	**-2,30**	**-0,18**	**-2,28**	174 / **92**	173 / **220**	Standard Error		YES
Rothfarb et al., 1987	0,05	0,24	**-0,30**	**-2,19**	NA	NA	No data in study		**0,22**	**2,08**	70 / **142**	49 / **126**			YES
Skoczylas, 1972	-0,05	-0,18	No data in study		0,13	0,46	**0,26**	**1,24**	**-0,68**	**-2,21**	25 / **25**	25 / **22**			YES
Summary (weighted)	**0,03**	**0,12**	**-0,05**	**-1,33**	**0,00**	**0,74**	**-0,07**	**-2,73**	**0,11**	**3,81**					
Summary Elem. Spanish (weighted)	**0,00**	**-0,29**	**-0,06**	**-1,41**	**-0,06**	**-1,28**	**-0,09**	**-2,93**	**0,11**	**3,74**					
# Elem. Span.	**7**		**4**		**6**		**6**		**5**						

Note: Shaded cells in summary data are statistically significant.

The Z score is not calculated from the effect size or any of the statistics that go into the effect size. The Z score is calculated from the probability of the F ratio or the t statistic or other tests of significance. It can be calculated in Excel[1] or obtained from a number of web sites. It is usually easier to calculate a confidence interval than a Z score, using the formula described above, and were it not for our desire to attempt to replicate Greene that is, in fact, what we would do.

Rather than replicating his inappropriate "All tests in English" column, we have inserted a column that consists of just the English language tests in these studies. Our summary effect sizes in Table 3 are an insignificant effect size of -.05 for English/language, a statistically significant *negative* effect size for reading of -.07, and a statistically significant *positive* effect size for math of .11. These are all small effects whether statistically significant or not. The same general results hold when only Spanish elementary programs are examined.

Random Assignment. Greene argues that random assignment studies are the best studies and so should be given more weight. With respect to internal validity that is, of course, true. One can be certain that the relationship between the independent variable and the dependent variable is not confounded by the assignment rule since it is random. If there is no random assignment, that is, if students are allowed to select themselves for a treatment or if someone else selects students for a treatment on the basis of characteristics that are correlated with the outcome, one must statistically control for those characteristics in order to isolate the effect of the treatment and one can never be certain the controls are sufficient.

Greene denotes six studies as having random assignment, but one of these is an error. The Rothfarb, et al. study is characterized by random assignment of schools, not students, to treatment and control groups. Indeed, Rothfarb et al. acknowledge this in conducting multiple regression analysis to control for the differences in student characteristics between schools. Excluding Rothfarb et al. leaves only four studies with random assignment. Of the four studies with random assignment, only two were of Spanish elementary programs.

---

[1] To calculate the two tailed probability of the F ratio in Excel: click on function, statistical, Fdist. In the popup table, X=f ratio, deg_freedom1=numerator df (between df of k-1), deg_freedom2=denominator df (within df of n-k) where n =total sample, and k=number of groups. The summary formula is FDIST(fratio,numdk,dendk). The convention in meta-analysis is that the Z score is calculated from a one-tailed probability since the Z score calculates the number of standard deviations from the mean, not conditioned on the direction, as if one knows which group will come out ahead. This is a questionable assumption, but we bow to convention on this issue. In order to obtain a one-tailed probability, the two-tailed probability is divided by 2. This means that a one-tailed probability will be smaller and since the probability is the probability that the relationship might have happened by chance, it is more likely that the difference between groups will be found to be statistically significant. To calculate the Z score from the one tailed probability in Excel: click on function, statistical, NORMSINV--in the popup window, insert the one-tailed probability if the experimental group is worse or 1 minus the one tailed probability if the experimental group is better**.** If the two-tailed probability has more than 5 zeros to the right of the decimal point, a .000001 will have to be added to the formula for the FDIST as in (FDIST(fratio,numdk,dendk))+.000001) or you can go to the web and find sites that will allow more than 5 zeros.

These two studies, Huzar and Plante, illustrate the problem with random assignment experiments—they all too often lack external validity or generalizability. The treatment programs in these two studies seem to have had the same amount of English instruction as the mainstream classroom and the students learned to read in English at the same time or before they learned to read in Spanish which is probably why the researchers or administrators could get away with random assignment without having a lawsuit on their hands. In short, these are not your typical Spanish bilingual education programs as Slavin and Cheung admit in the conclusion of their paper, but Greene ignores.

The secondary programs with random assignment (Covey and Kaufman), in particular, seem to have had little Spanish language instruction and may have consisted only of after-school tutoring by Spanish speaking aides. Since we no longer accept secondary bilingual education programs, the lack of external validity of these two studies is a moot point for us.

Reanalyzing Slavin and Cheung's 2004 Sample and 2005 Table

There are many formulas for computing an effect size although the two most common seem to be Cohen's d and Hedge's g. Slavin and Cohen used Cohen's d for their effect sizes in the July 2004 paper. This effect size has the control group's standard deviation in the denominator rather than the pooled standard deviation as is the case with Hedge's g. Since the treatment and control groups in these studies occasionally had very different Ns, we would recommend Hedge's g over Cohen's d. In fact, Slavin and Cheung have now come to this conclusion and Cheung has sent us a revised table which now has Slavin's g.

Appendix 4 shows the studies that Slavin and Cheung (2004) included and rejected in their meta-analysis which is adapted from their Appendix 1 with columns added by the first author of this paper, noted as CR. The first column added labeled "CR Comments on Source" shows the studies that Greene accepted as well as errors that Slavin and Cheung made in attributing the citation for a study. As noted above, there is disagreement between Greene and Slavin and Cheung with regard to criteria. Whereas Greene accepted studies where the pretest occurred after treatment was underway (as did Rossell and Baker), Slavin and Cheung did not. Greene, on the other hand, rejected studies where the only control variable for the differences between groups was a pretest, but Slavin and Cheung accepted those studies (as did Rossell and Baker).

As a result of these differences in criteria and other issues, Slavin and Cheung accepted Alvarez (1975), but Greene rejected it because he believed it inadequately controlled for differences between bilingual and English-only students (i.e. the only control was a pretest). As shown in the final column, however, the Alvarez study violates Slavin and Cheung's criterion that the pretest had to be given before the treatment was under way. Indeed, of the 16 elementary reading programs that Slavin and Cheung accepted, five violated their criterion that pretests had to be administered before treatments were underway. In short,

Slavin and Cheung were inconsistent, although in our opinion it is probably not possible to be entirely consistent with these messy, complicated studies.

Appendix 5 shows our replication of Slavin and Cheung's Cohen's d. It is an adaptation of their Table 1 from their July 2004 paper. Appendix 6 shows our replication of their revised Table 1 now using Hedge's g as the effect size, emailed to the second author on February 5, 2005. Slavin and Cheung did not report summary results or even sub-category results in their tables, although that is presumably a major advantage of meta-analysis over the vote count method. They also did not report significance levels or confidence intervals for any of their studies.

The sample sizes that Slavin and Cheung report match those that we found for most of the studies. The only real discrepancy was with the Campeau study of Houston in which Slavin and Cheung reported the sample size for one of the grades for one of the years, while we report the sample size for the last year for all groups. Slavin and Cheung also report only two cohorts for Cohen (1975) when in fact there are three. Their numbers for Kaufman come from the initial sample. The sample of students who actually took the post-test is the number we believe should be reported. We have no idea where the sample N for Covey comes from.

We object to including the Maldonado (1994) study. The effect size of 2.21 with Cohen's d and 1.66 with Hedge's g (we got 1.73), are unbelievable. Effects this large are just not obtained from educational treatments so there is something else going on. As described in the study, the educational treatment is not only a double dose of reading which the control group did not get, but other treatments not received by the control group. One of the more important of these other treatments is that the teacher assigned to the treatment group had experience working with "integrated bilingual special education" and teaching bilingual students with learning disabilities. The control group teacher apparently had no experience working with bilingual students with learning disabilities. The teaching strategies used by the experimental group teacher include a wide range of strategies beyond the language of instruction. The control group program is hardly described at all except to say that some of the strategies were the same for both groups. Because this study had random assignment (of students, not teachers), there were no statistical controls for any of the other characteristics of these two programs or students.

Indeed, the results are so unbelievable as to make one wonder if the problem extends beyond the fact that the experimental group had an experienced teacher who used a wide range of strategies in addition to changing the language. Not only did the treatment group have an astonishing 29 point gain in their CTBS reading scores, but the control group actually had a nine point decline in achievement. Neither effect is credible even if the treatment group received significantly better instruction and one can only wonder if the researcher made a mathematical or other kind of error. For all of these reasons, including the fact that this is a study of special education students, we exclude this study. Even if the data were believable, the study has limited generalizability.

In addition, we exclude all but the Corpus Christi study of Campeau (per Rossell and Baker), the only one that seems to have a treatment and a control group and some statistical control for pretreatment differences. The effect sizes that Slavin and Cheung report for the Campeau et al. study of Santa Fe are problematic as there is not enough information in that study to create an effect size. Cheung is still struggling with the issue of exactly how to estimate an effect size for this study since there is no data.[2] We think no effect size can be created from this study without literally making up data and so we have left the cells empty.

It is curious that the Slavin and Cheung review left the following studies (J. R. Maldonado 1977; Cohen 1975; Alvarez 1975; Ramirez, et al. 1991; and Kaufman 1968) non-quantified in the Cohen's d analyses or arbitrarily assigned them an effect size of zero in the Hedge's g analyses (J. R. Maldonado 1977; Cohen 1975; Ramirez, et al. 1991) when in fact they *do* have enough data to compute an effect size. Greene also computed effect sizes for Ramirez and Kaufman (but not the others as they were rejected or not considered).

We calculated our own Cohen's d and Hedge's g effect sizes for these studies and measured the significance of the effects using the 95% confidence interval. If the interval does not include 0, the effect size is statistically significant. Of the 12 Cohen's d effect sizes Slavin and Cheung calculated, seven were significant and five were not. Of the 14 Cohen's d effect sizes we calculated, five were significant and nine were not. Of the 18 Hedge's g effect size Slavin and Cheung calculated or arbitrarily assigned a zero to, seven were significant and 11 were not. Of the 14 Hedge's g effect sizes we calculated, four were significant and 10 were not.

Table 4 contains the summary statistics from Appendix 5 and 6. The average weighted Cohen's d effect size across all of Slavin and Cheung's studies, using our effect size where they had none, and their effect sizes for the other studies is .34, small but statistically significant. Only 44 percent of the studies had a significant effect size. Across just the studies where they calculated an effect size, it is .57, medium and statistically significant. Only 58 percent of the studies had a significant effect size. Our Cohen's d effect size for the Spanish elementary bilingual education programs, excluding the Campeau and Maldonado studies, is .14, but still (barely) statistically significant. Across all Spanish elementary bilingual education programs, only 36 percent of the studies had significant effect sizes.

[2] Email communication with second author, 12/8/04 and 2/12/05.

Table 4: A Comparison of Summary Effect Sizes by Slavin & Cheung and Rossell & Kuder

	Slavin & Cheung			Rossell & Kuder	
	All Studies*	**All Studies with S&C ES****	**Stat. Sig.**	**Spanish Elementary**	**Stat. Sig.**
COHEN'S d					
Effect Size	**0,34**	**0,57**	**Yes**	**0,14**	**Yes**
Lower C.I.	0,26	0,45		0,03	
Upper C.I.	0,43	0,73		0,26	
% studies statistically significant	44%	58%		36%	
N in Analysis	18	12		14	
HEDGE'S g					
Effect Size	**0,25**		**Yes**	**0,10**	**No**
Lower C.I.	0,17			-0,01	
Upper C.I.	0,34			0,22	
% studies statistically significant	39%			29%	
N in Analysis	18			14	

* Includes Cohen's d effect sizes calculated by Rossell & Kuder if Slavin and Cheung did not report them.

** Only includes studies that Slavin & Cheung computed an ES for.

The average weighted Hedge's g effect size for Slavin and Cheung across all studies, including the arbitrary zero effect sizes assigned to some studies, is .25, small but statistically significant. Only 39 percent of the students had significant effect sizes. Our Hedge's g effect size for the Spanish elementary bilingual education programs, excluding the Campeau and Maldonado studies, is .10, not statistically significant. However, 29 percent of the studies had significant effect sizes.

But it must be emphasized that most of these studies were not of conventional bilingual education programs as Slavin and Cheung admit at the end of their paper. As noted above, the students received a double dose of reading (hence the term paired bilingual), one period in Spanish and one period in English and in several programs had no less English instruction than students in the mainstream classroom. The theory underlying bilingual education in the U.S. is that one must learn to read and write first in the native tongue and receive

subject matter in the native tongue *before* transitioning to English. These programs violate that theory.

What Slavin and Cheung do not consider in their paper, although Slavin admitted this in personal communication to the first author in Berlin, is the possibility that the effect on English language achievement is of the double period of reading, not the language of instruction. Indeed, it is very possible that if the double period of reading had been in *English*, the effect might be even more positive than they found in their sample and might be positive rather than the no effect we found. At this point, we can say that our reanalyses of both Greene (1998) and Slavin and Cheung (2004, 2005) do not support the conclusions they draw regarding the superiority of bilingual education over a mainstream classroom.

Reanalyzing Rossell and Baker

Table 1b and Appendix 1b show a revised vote count tally based on our new criterion—no programs of less than a school year, no secondary programs, and no non-Spanish speaking bilingual education programs. The two studies that are actually redundant (Ariza 1988 and Curiel, Stenning, and Cooper 1980) have also been removed. We also recategorized two studies. The El Paso studies have been moved from the category of TBE versus mainstream classroom to TBE versus structured immersion. Gersten (1985) has been moved from TBE versus structured immersion to structured immersion versus ESL (the program that had been called bilingual education). The studies that have been removed or relocated are crossed out and those that were inserted in a new place are bolded and underlined.

As can be seen, this does not change our findings in any important way. The percentages vary only slightly. On average, the best program is structured immersion and the more native tongue instruction, the lower one's achievement in the second language. Nevertheless, there are enough exceptions to this overall finding that it is possible to also say that a little bit of native tongue instruction does not hurt and might help if the native tongue is Spanish. We maintain, however, that this is more consistent with programs that we call structured immersion, or sheltered English immersion in the U.S., than it is with transitional bilingual education as described in the literature—a program where children must learn to read and write in their native tongue initially and must reach literacy in the native tongue before being transitioned to English.

Table 1b: Revised % of Methodologically Acceptable Studies With Program Superiority, Equality, or Inferiority by Achievement Test Outcome*

	READING**	LANGUAGE	MATH
TBE v. Submersion (Mainstream)			
TBE Better	20%	14%	10%
No Difference	51%	29%	55%
TBE Worse	29%	57%	35%
Total N	35	7	20
TBE v. ESL Pullout			
TBE Better	0%	0%	20%
No Difference	50%	50%	40%
TBE Worse	50%	50%	40%
Total N	4	4	5
TBE v. Mainstream/ESL			
TBE Better	18%	9%	12%
No Difference	51%	36%	52%
TBE Worse	31%	55%	36%
Total N	39	11	25
TBE v. Structured Immersion			
TBE Better	0%	0%	0%
No Difference	14%	25%	50%
TBE Worse	86%	75%	50%
Total N	14	4	10
Structured Immersion v. ESL			
Immersion Better	100%	0%	100%
No Difference	0%	0%	0%
Total N	4	0	1
TBE v. Maint. BE			
TBE Better	0%	0%	0%
Total N	0	0	0

* Studies are listed in more than one category if there were different effects for different grades or chorts.

** Oral English achievement for preschool programs.

Original Source: C. Rossell and K. Baker, "The Educational Effectiveness of Bilingual Education,"

Research in theTeaching of English, 30 (1), February 1996: 1-74.

Testing Rates

None of the reviews, including Rossell and Baker, controlled for the considerable difference in testing rates between Spanish speakers in bilingual education and those in all-English classrooms. There is a consistent bias in virtually all evaluations that compare Spanish bilingual education programs in the U.S. to an alternative program. Teachers can decide when their English Learners are ready to take standardized achievement tests. Teachers in bilingual education program test their English Learners at lower rates than do teachers in all-English programs because they believe that it is unreasonable to administer English language tests to students who are learning literacy in their native tongue. However, this gives the bilingual education programs an unfair advantage over all-English programs because a much larger number of low achieving students will not be included in the evaluation of the bilingual education program than is the case with the all-English program. It is the lowest scoring students who are deemed not ready to be tested.

Individual student data from California and the U.S. show even more striking disparities in testing rates. Bali (2000) has obtained individual student data and program testing rates pre and post Proposition 227 for Pasadena Unified in southern California. She found a 50 percent testing rate for the English Learners in bilingual education in Pasadena in 1997-98, but an 89 percent testing rate for the English Learners in ESL in the same district.

Similar disparities in testing rates were found in the Los Angeles Unified School District in 1996-97. The school district's report showed English Learners who were in bilingual education for five years outscored English Learners in all-English classes on the Stanford 9. However, only 61 percent of the students in the bilingual program were thought to know enough English after five years to be able to take the test, but 97 percent of the students in the English language program took the test (Los Angeles Unified 1998). This 37 point differential is very close to the 39 point differential Bali found in Pasadena.

Similar disparities can be found in the Ramirez et al. (1991) nationwide study of more than 1,000 children in 9 school districts, 46 schools, and 136 classrooms across 5 grades. Eighty-nine percent of the structured immersion students were tested in K-1, but only 61 percent of the early exit bilingual education students were tested. In grades 1-3, 42 percent of the structured immersion students were tested, but only 29 percent of the early exit bilingual education students were tested. The Ramirez study found no difference between the two programs, but this underestimates the benefit of immersion and overestimates the benefit of bilingual education since far fewer students were tested in the bilingual program.

The first author has done similar analyses of testing rates in California (Rossell, 2002; 2003). The higher the percentage enrolled in bilingual education, the lower the testing rate. Thus, the evaluations of Spanish bilingual education in the U.S. are biased by the fact that only the best students are tested in Spanish bilingual education programs, but almost all English language learners are tested who are in a mainstream classroom. In addition, these testing rates can be thought of as outcomes. If there are more ELLs in bilingual education deemed not ready to be tested than in the mainstream classroom or structured immersion,

even after several years in the program, then the bilingual education program is less effective than the alternative in teaching the language that will appear on the test.

Evaluating Bilingual Education in California

In June 1998, California voters voted to make the default assignment for English language learners a structured immersion classroom. Before that it had been bilingual education.

Table 5.1 of Rossell (2002) is a regression equation predicting the effect of the percentage of English language learners enrolled in bilingual education on an elementary school's 2001 reading and math test scores[3] controlling for their 1998 test score and their percentage poor in 2001 (enrolled in Calworks, the state poverty program).[4] The 1998 test score is basically a control for the characteristics of the school that are not captured in the poverty rate.[5] The test scores for ELLs are low (on a scale from 0 to 100), but that is because they are supposed to be low - an English language learner is a student who scores low in English. This also means there is a ceiling on how much progress can be made in ELL test scores. This is because when ELL scores get above a certain level (around the 36th to 50th percentile depending on the district), they no longer appear in the English language learner category. That category is *only* of low scorers.

The regression analysis indicates that the percentage enrolled in bilingual education is significantly and negatively related to a school's test score in both reading and math even after controlling for poverty rates and initial test scores. If we solve the equation for 100, 50, and 0 percent of a school's English Learners in bilingual education in 2001, an elementary school's reading score is increased by six points in reading and three points in math if they have no bilingual education enrollment compared to a school that has all its English Learners enrolled in bilingual education.

This analysis may not show the true effect of bilingual education, or its inverse, English language instruction, on school achievement since it appears that bilingual education in California has been changed by Proposition 227 - more English is being used - and because all but a handful of schools reduced their bilingual education enrollment even if they did not eliminate it entirely. Trying to isolate the true effect of a program that is no longer the same or the true effect of sheltered English immersion when it also had an effect on other

---

[3] This is the school's average NCE converted to a national percentile rank. The state does this conversion.

[4] The percentage of English Learners tested in reading or math was not significant at the school level and is not shown. It may be that in a statistical analysis at the school level, the problem of countervailing tendencies - low test rates occur in schools with low achievement - muddles the advantage of not testing the very lowest scoring students. Because the higher scoring schools test more of their students, the sign for the testing rate variable is positive, although insignificant.

[5] The state data also include the achievement of all students in a school, but that is not a good control variable since the English Learners comprise a large percentage of all students in the schools that formerly had bilingual education programs. In addition, most of the fluent English proficient (FEP) students were once English Learners and so controlling for the achievement gains of fluent English proficient students wipes out part of the treatment effect for English Learners.

programs is a difficult task even at the individual level and it is even more difficult at the school level.

Moreover, as noted above there is a ceiling effect that is present in the state data since it is not possible to examine the achievement of redesignated English language learners. In order to know the true effect of Proposition 227 or the remaining bilingual education programs, one must be able to follow English Learners after they are redesignated fluent English proficient and unfortunately, at this point in time that is not be possible with school level data.

Individual student data still suffers from the testing rate bias favoring bilingual education, but at least it is possible to determine the program the student is enrolled in. Bali (2000) has analyzed the achievement of individual English Learners in the Pasadena Unified School District using data provided by them. In 1998, 53 percent of Pasadena's English Learners were enrolled in bilingual education. After Proposition 227, less than two percent of English Learners were enrolled in bilingual education. Bali used the Heckman (1979) selection model to control for the selection bias introduced by the lower testing rate for the bilingual education program in 1997-98.

The effect of being in a bilingual education program in 1998 is negative and statistically significant, but the magnitude was only 2.4 points in reading and a half point in math. The effect of putting these same English Learners in a structured immersion classroom the next year was to eliminate the small gap between English Learners who had been in bilingual education and those not in bilingual education.

These findings are not that different from what I obtained in a school achievement analysis. *School* achievement in reading increases by six points if all children are enrolled in bilingual education compared to a school where none are. School achievement only increases by three points in math if all children are enrolled in bilingual education compared to a school where none are.

Conclusion

The best approach to educating second language learners is not an issue that can be solved by meta analysis and probably not by any other statistical approach. There is too much disagreement over what constitutes scientific research and too little scientific research. None of the research is perfect and much of it is extraordinarily complicated with many, many analyses and outcomes. Honest and competent professionals can legitimately disagree as to whether a study is good enough to be relied upon.

Nevertheless, we are confident that structured immersion is the best approach to educating second language learners and that most second language learners should not be in that protected environment for longer than a year or two. Virtually all of the comparisons that Greene (1998) and Slavin and Cheung (2003) made were of bilingual education and a mainstream classroom. Much of what is valuable about bilingual education (the sheltered envi-

ronment, the caring and trained teacher, the engagement of students in instruction they can understand, and the use of the native tongue to clarify when necessary or possible) can be obtained in a structured immersion classroom without the reduction in the second language that can have negative consequences on a child's achievement in the second language.

On the other hand, Spanish bilingual education in the U.S. is not a disaster and children do learn English. Rossell and Baker (1996b) hypothesized that if Spanish reading was taught briefly when a child literally knows no English, it might be a superior approach for teaching reading and simple math to Spanish speakers. The problem was the theory that Spanish must be mastered before English. That all too often keeps children in Spanish too long and reduces their English language achievement. But we believe that programs that use only some native tongue in the beginning are closer to structured immersion than they are to the bilingual education that is described and supported in the literature.

References

Baker, Keith and deKanter, Adriana. 1981. The effectiveness of bilingual education programs: A review of the literature. Final draft report. Washington, DC: U.S. Department of Education.

Baker, Keith and deKanter, Adriana. 1983. "Federal Policy and the Effectiveness of Bilingual Education." In Keith A. Baker, Adriana A. deKanter, eds., Bilingual Education. Lexington, MA: D.C. Heath and Company.

Baker, Keith. 1987. "Comment on Willig's `A Meta-Analysis of Selected Studies in the Effectiveness of Bilingual Education'." Review of Educational Research. 57(3):351-362.

Bali, Valentina. 2000. "'Sink or Swim': What Happened to California's Bilingual Students After Proposition 227?" Unpublished paper, Pasadena, CA: California Institute of Technology.

Bali, Valentina. 2001. "'Sink or Swim': What Happened to California's Bilingual Students After Proposition 227?" State Politics and Policy Quarterly, 1(3): 295-317.

Bruck, Margaret, Jola Jakimik, and G. Richard Tucker. 1971. "Are French Immersion Programs Suitable for Working-Class Children? A Follow-up Investigation." Word 27:311-341.

Cohen, Jacob. 1988. Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press.

Cooper, H.M. 1989. Integrating Research: a Guide for Literature Reviews. (2nd ed)., Newbury Park, Calif.: Sage Publications.

Cziko, Gary A. 1975. "The Effects of Different French Immersion Programs on the Language and Academic Skills of Children from Various Socioeconomic Backgrounds". M.A. thesis, McGill University.

Genesee, Fred. 1976. "The Suitability of Immersion Programs for all Children." Canadian Modern Language Review 32:494-515.

Gersten, Russell, Baker, Scott and Otterstedt, Janet. 1998. "Further Analysis of: A meta-analysis of the effectiveness of bilingual education, by J.P. Greene (1998)," Eugene, OR: Eugene Research Institute.

Greene, Jay P. 1998. "A Meta-Analysis of the Effectiveness of Bilingual Education," Unpublished paper of the Tomas Rivera Policy Institute, Claremont Graduate School.

Greene, Jay P. 1997. "A Meta-Analysis of The Rossell And Baker Review Of Bilingual Education Research, 21 (2 and 3), Bilingual Research Journal, Spring and Summer.

Heckman, J.J. 1979. Sample Selection Bias as a Specification Error, Econometrica, 47, January, 153-161.

Lipsey, Mark W. and Wilson, David B. 2001. Practical Meta-Analysis. Newbury Park, Calif.: Sage Publications.

Los Angeles Unified School District. 1998. "Clarification of English Academic Testing Results for Spanish-Speaking LEP Fifth Graders."

Meyer, Michael M. and Feinberg, Stephen E. (Eds.) 1992. Assessing Evaluation Studies: the Case of Bilingual Education Strategies. Washington, D.C.: National Academy Press.

Rosenthal, R. 1991. Meta-analytic procedures for social research. Newbury Park, Calif.: Sage Publications.

Rossell, Christine. 2003. "The Near End of Bilingual Education," Education Next, vol. 3 (4), Fall 2003: 44-52.

Rossell, Christine. 2002. "Dismantling Bilingual Education, Implementing English Immersion: the California Initiative," February 20.

Rossell, Christine. 1980. "Social Science Research in Educational Equity Cases: a Critical Review," Review of Research in Education, 8, 237-295.

Rossell, Christine and Ross, J. Michael. 1986. "The social science evidence on bilingual education." Journal of Law and Education 15:385-419.

Rossell, Christine and Keith Baker. 1996a. "The Educational Effectiveness of Bilingual Education," Research in the Teaching of English, February, 30 (1): 7-74.

Rossell, Christine and Keith Baker. 1996b. Bilingual Education in Massachusetts: the Emperor Has No Clothes. Boston, MA: Pioneer Institute.

Shadish, William R. and Haddock, C. Keith. 1994. "Combining Estimates of Effect Sizes." In Harris Cooper and Larry V. Hedges, Eds. The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Slavin, Robert E. and Alan Cheung. 2003 (December). "Effective Reading Programs for English Language Learners: A Best Evidence Synthesis." Report 66. Washington, D.C.: Center for Research on the Education of Students Placed At Risk. Available at http://www.csos.jhu.edu/crespar/reports.htm.

Slavin, Robert E. and Alan Cheung. 2004 (July). "Synthesis of Research on Language of Reading Instruction for English Language Learners." Paper presented at the Workshop on "The Effectiveness of Bilingual School Programs for Immigrant Children" at the Social Science Research Center Berlin (WZB), Programme on Intercultural Conflicts and Societal Integration (AKI),Nov. 18-19, 2004.

Tucker, G. Richard., Wallace E. Lambert, and Alix d'Anglejan. 1973. "French Immersion Programs: A Pilot Investigation." Language Sciences 25:19-26.

Willig, Ann C. 1985. "A Meta-Analysis of Selected Studies on the Effectiveness of Bilingual Education." Review of Educational Research. 55(3):269-317.

Wolf, Fredric M. 1986. Meta-Analysis: Quantitative Methods for Research Synthesis. Newbury Park, CA: Sage Publications.

From Cure to Curse: The Rise and Fall of Bilingual Education Programs in the Netherlands

Geert Driessen

For about 35 years, two types of bilingual education programs have been offered to ethnic minorities in the Netherlands. The first are the bilingual reception models, which have been applied only in a few schools. The second, Mother Tongue Instruction (MTI), was more widespread and took up as much as 10% of the time spent at school. The focus in this paper is on MTI. As MTI has been abolished as of 2004, this chapter bears the character of an evaluation of the policy developments, their implementation, and effects. The following text provides an overview of policy developments with regard to MTI. It takes a closer look at the arguments on which that policy was based and presents arguments held against this policy. Subsequent paragraphs outline major results of Dutch evaluation studies of bilingual education. The paper concludes with some observations about why, given the unfavorable framework, bilingual education in the Netherlands could not become a success story.

1. Ethnic Minorities in the Netherlands

Since World War II, various groups of immigrants have come to the Netherlands, mainly for political and economic reasons. They comprise immigrants from the former Dutch colonies, i.e. the Dutch East Indies (largely Indo-Europeans), the Moluccas, Surinam, and the Netherlands Antilles. As a result of their ties with the (former) motherland, these immigrants were already somewhat acquainted with the Dutch language and culture.

Primarily during the early 1960s, so-called foreign or guest workers arrived from the Mediterranean countries, e.g. Italy, Spain, Portugal, Greece, Yugoslavia, Turkey, and Morocco. Due to family reunification and the immigration of spouses, there has been a continuous influx of immigrants ever since, particularly coming from Turkey and Morocco. One characteristic these immigrants have in common is their low level of education. Furthermore, their languages and cultures differ considerably from the Dutch. This is true particularly for the Turks and Moroccans.

Refugees and asylum seekers who arrived from Eastern Europe, Latin-America, Asia, Africa, and the Middle-East, are a very diverse and ever-changing group, both in terms of languages and culture.

Furthermore, there are immigrants from Western countries, such as Belgium, Germany, the UK, and the USA, who usually have a middle or higher socio-economic status.

Only some immigrant groups are the targets of an ethnic minority policy. Immigrants from the Dutch East Indies are excluded from this policy as are immigrants from Western countries. In essence the ethnic minority policy is aimed at those immigrants (and their descendents) with a low socio-economic status. Depending on which criterion is applied, the percentage of people resident in the Netherlands who belong to ethnic minorities (henceforth: minorities) varies from 7 to 19 per cent. Based on the 'country of birth' (of the individuals and their parents), in 2003 the largest minority groups in the Netherlands were of Turkish (341,000), Surinamese (321,000), Moroccan (295,000), and Antillean (129,000) origin, out of a total Dutch population of 16 million (Tweede Kamer 2003). A large share of the minority population lives in the major cities. On average they are relatively young. As a consequence, in cities such as Amsterdam, Rotterdam, Utrecht, and The Hague more than half of the primary school pupils are members of minority groups.

2. Educational Policy Relating to Minorities

In the Netherlands, educational policies relating to minorities have a relatively short history and have repeatedly been revised. Four distinct phases can be distinguished (Driessen 2000; Eldering 1989; Fase 1994; van het Loo, de Spiegeleire, Lindstrom, Kahan & Vernez 2001):

1. The period before 1980 was characterized by a so-called two-track approach. The government assumed that the children of guest workers were going to stay only temporarily. Consequently, all efforts were directed toward the integration of children into the Dutch education system and, simultaneously, toward preparing them for their return to their home country. In the classroom, this meant that the children received instruction in their home language (bilingual education) as well as extra tuition in Dutch (Dutch as a second language).
2. Around 1980 the government abandoned the idea that the immigrants would return to their home countries. From then on, education aimed at preparing immigrants for their role in Dutch society. The objective was to enable them to become full members of society with respect to socio-economic, social, and democratic aspects, while retaining their own cultural background. A number of educational policies were designed which were specifically targeted at cultural minority groups. Among other things, this meant that schools with immigrant children were given additional resources. In addition to minority policies, a separate Social Priority Policy program aimed at improving the opportunities of native Dutch working-class children.
3. During the third phase, beginning in 1985, the Educational Priority Policy program (EPP) came into effect. This program integrated policies relating to minorities and the Social Priority Policy program. The underlying idea was that immigrant children and Dutch working-class children suffer from comparable disadvantages and that the causes of their disadvantages are very similar. The aim of the EPP was to eliminate or reduce such disadvantages.

4. At the beginning of the 1990s, it became clear that although in a broad sense some progress in the educational position of minority children had been achieved, their performance still lagged far behind that of native Dutch children. As a consequence, the government opted for a new Educational Disadvantage Policy (EDP). Its key concepts were decentralization, deregulation, and increased autonomy for the municipalities. The EDP came into effect in 1998 but has already been revised. An important alteration concerns the shift in responsibility from the municipalities to the boards of governors of schools.

3. Bilingual Models and Mother Tongue Instruction: Policies from 1970 to 2004

Bilingual education was one of the elements of the Educational Priority Policy targeted at minorities. In Dutch primary education, two models of bilingual education can be distinguished: bilingual reception models and Mother Tongue Instruction.

The **bilingual reception models** can be further subdivided into transitional models and simultaneous models (Baker 1988). Both were applied in the lower grades of primary school. In the transitional model, the children received instruction in the minority language in the third grade[1], and mixed instruction in the minority language and in Dutch in the fourth grade. From the fifth grade onwards they received Dutch instruction only. In the simultaneous model, the children for two or three years received roughly half of their instruction in the minority language and the other half in Dutch. The purpose of these types of models is twofold: On the one hand, it is expected that, with the aid of the minority language, the children will be better able to master the Dutch language; on the other hand, they are expected to attain a certain level of functional literacy in the minority language. These models were only applied in a few schools in the Netherlands, mainly in the 1980s. The languages concerned were Turkish and Arabic.

Mother tongue instruction (MTI)[2] was first introduced in 1967. Initially it was organized and financed by the immigrant parents themselves with or without the support of their embassies. From 1970, the Dutch government partly financed MTI. The aim of this form of MTI, which was not part of regular education, was to ease re-integration upon their return to the country of origin. After 1974 a two-fold policy in relation to the immigrants was pursued: it was assumed that many immigrants would continue to stay in the Netherlands, while the remigration idea was not given up. As a result, the aims of MTI were twofold: For future re-migrants it was considered essential to enable them to fit into the education

[1] Dutch primary schools are attended by 4 to 12-year-olds. In grades 1 and 2 play takes up a central place. In grade 3 formal instruction in reading, maths and writing starts. After the last year, grade 8, pupils move on to secondary school which they attend for 4 to 6 years.

[2] In the Netherlands this was referred to as 'Onderwijs in Eigen Taal en Cultuur' (OETC), which literally stands for 'Teaching in the Own Language and Culture'. Later the cultural component was scrapped and it became OET. Still later it was referred to as 'Onderwijs in Allochtone Levende Talen' (OALT) which stands for 'Teaching in Non-Indigenous Living Languages'.

systems of their countries of origin. For those staying, MTI was meant to promote integration into Dutch society. With regard to integration, the idea of strengthening minority identities became more important. From 1974 onwards, primary schools were offered the possibility of incorporating MTI into their regular program. MTI was not restricted to language instruction, but included a cultural component (e.g. aspects of the history, geography, and culture of the country of origin).

Around 1980, the government discarded the idea that the immigrants' stay was only temporary and accepted that they were to stay for good. In a policy document published in 1981 three functions were attributed to MTI:

a) to contribute to the development of the self-concept and the self-awareness of students;

b) to help preserve the possibility of maintaining contact with family members and friends in the country of origin;

c) to ease the reintegration into the education system of the country of origin in case of return.

In 1983 another policy document on MTI was issued. Here the government suggested that MTI be integrated into regular education. The functions ascribed to MTI in this policy document deviated somewhat from the ones mentioned in the 1981 one. These were as follows:

a) the development of a positive self-concept and self-awareness;

b) to close the gap between school and home environment;

c) to make a contribution to inter-cultural education (intended for all pupils).

Officially at least this meant that the aim of reintegration was abandoned. Instead MTI from then on aimed at the acculturation of pupils in the Netherlands who did not speak Dutch. Furthermore, the document contained a plea for a strong role of language instruction, whereas cultural instruction was transferred to the so-called inter-cultural education (which was intended for both the native Dutch and the immigrant children). It was also argued that MTI should contribute to achieving the general aims of educational policy with regard to minorities.

1984 marked an important change for MTI, as it then acquired a legal basis. Meanwhile, MTI was increasingly regarded as a means of promoting the educational opportunities of immigrant pupils. This role of MTI was strengthened by the fact that from 1986 onwards it was regarded as part of the new Educational Priority Policy (EPP), i.e. the policy for native-Dutch as well as immigrant pupils, aimed at combating educational disadvantages resulting from social, economic or cultural factors. During this period, the idea that a command of the minority language was beneficial for the learning of Dutch increasingly gained ground. Furthermore, it was also widely assumed that the teaching of the minority language would signify recognition and appreciation of the cultural identities of the groups involved and

that this would promote their emancipation. This type of MTI can probably best be characterized as a restricted maintenance model – a bilingual education model which aimed at full proficiency in Dutch and only limited proficiency in the mother tongue.

In 1991 another policy document on MTI was published. Now the emphasis was even more on linguistic arguments and the cultural component was scrapped completely. The most important proposition of this document was that minority-language instruction should be integrated into the overall language policy of the school. Its main function should be to support the acquisition of Dutch and academic advancement in general. In addition, however, MTI was meant to give access to the home culture and cultural heritage and thus promote the development of self-awareness. Greater use of the minority language within the framework of bilingual education in the first three to four grades of primary school was encouraged. In those cases in which the home language deviated from the official language of the country of origin, the home language was to be used. But from the fourth up to the eighth grade of primary education, instruction was to be given as a separate subject now mainly in the official language of the country of origin. It was suggested that MTI should be integrated into the normal school curriculum and general instruction in their own school as much as possible (children could also attend these lessons in other schools than their own, which was the case when there were only few children at a school).

In the second half of the 1990s, it became clear that the disadvantages suffered by minorities had not diminished over the years: minority students were already considerably behind when they entered primary school, and they did not catch up throughout the course of their school career. The aim of the EPP had apparently not been achieved. As far as bilingual education was concerned, it should be noted that the reception models had only been implemented at a small number of schools. Opinions varied on the effects of these experiments. This also applied to MTI. Furthermore, there was a great deal of discussion about the extent to which this kind of education was (still) useful and whether it did not in fact reduce educational opportunities and result in segregation instead of integration. As will be explained in more detail below, it had not improved performance in the regular subjects, and for half of the minority children, Moroccan children in particular, positive effects in terms of their 'own' language and culture were practically nonexistent. In response to these developments the Ministry of Education published a new document concerning the future of MTI. According to the ministry there were three reasons for continuing MTI:

(a) it can enhance the emancipation and participation of minorities in Dutch society;

(b) it can be beneficial for the Dutch economy;

(c) it can help the interculturalization of Dutch society as a whole.

The ministry distinguished two types of MTI. In the first grades the 'own' language should function as a support for learning the Dutch language. In the upper grades MTI had an autonomous function as a form of cultural education. Conceived in this way, MTI was not connected with the regular curriculum and therefore to be offered after school time. An

important new aspect was that, although the ministry claimed to still support this after school MTI, it delegated the responsibility for the implementation to the local authorities.

After '9/11', the political climate changed dramatically in the Netherlands. Elections held in 2002 were completely dominated by the populist right-wing politician Pim Fortuyn, whose main theme was immigration and integration. As he claimed, many immigrants did not speak Dutch, even after having been in the Netherlands for decades, many were unemployed, and often they did not adhere to Western and liberal norms and values such as the emancipation of women and the separation of church and state. The success of Fortuyn (assassinated a few days before the 2002 elections) and his party is to be seen against the background of a growing resentment towards immigrants. Calls for assimilation instead of integration and maintenance of the minority languages and cultures have become influential (cf. Vermeulen & Penninx 2000). Now, in the first half of 2004, the Netherlands is being governed by a coalition of Christian-Democrats, Liberals, and Center-Democrats. These parties have adopted many of Fortuyn's right-wing ideas on immigration. They include the abolition of MTI as of 2004. According to the Ministry of Education priority should be given to the learning of Dutch. Another important change is that people who want to immigrate to the Netherlands have to learn Dutch in their native country and pay for these courses. If they do not pass the examination, they will not be allowed to settle in the Netherlands.

While only a limited number of children were involved in the bilingual reception models, all children of guest workers had, before 2004, been in principle eligible for the official form of MTI. In addition, Moluccan children and children of officially recognized refugees could also attend MTI. Surinamese and Antillean children, however, were excluded from MTI because the official language of these former colonies is Dutch. Chinese MTI was not subsidized either, because Chinese immigrants are not included in the official minority policy.[3] Under the EPP, MTI could be attended for a maximum of 2.5 hours a week during school hours, and for 2.5 hours after school. MTI was open to first- and second-generation immigrant children. The children were to be taught in the official language (the standard language) of their (or their parents') native country. In 1995, 67,000 children enrolled in MTI. Of these, 61,000 were of Turkish or Moroccan origin, which represented 73 per cent of the total number of Turkish and Moroccan children in primary education. It was estimated that in 1998 participation dropped to 67 per cent and in 2000 to 57 per cent. In total, a full-time equivalent of approximately 1000 teachers were appointed to give instruction at 1,200 schools.

[3] Very little is known about the unofficial form of MTI which is not subsidized and organized by the government, but by minority parents themselves (e.g., Chinese MTI). By definition this variant takes place after school hours and the number of hours is, in principle, unlimited.

4. The Pros and Cons of Mother Tongue Instruction

The previous paragraphs presented an overview of the aims of MTI as pursued by the Dutch government. These official aims were founded on an extensive range of arguments in favor of MTI (cf. Demirbaş 1990; Driessen 1990; Extra, Folmer & van der Heijden 1992; Fase 1987). However, there was also opposition to MTI, and the following overview lists some of the main arguments in favor of and against MTI.

Educational psychology arguments

For

- Arguments in favor of MTI are often based on the interdependency and threshold hypotheses of Jim Cummins. Broadly summarized, these two hypotheses assume that the language proficiency level in Dutch depends on the proficiency level in the mother tongue and that a certain level has to be attained in both languages before bilingualism can be assumed to have positive effects on cognitive development. Important conditions for success are, first, adequate exposure to the second language (either in school or the environment) and, second, adequate motivation to learn the second language. Insufficient development of the home language carries the risk of semi-lingualism, i.e. the child does not speak either language very well.

Against

- Both hypotheses are very abstract and have not really been described in operational terms (cf. Cummins 1991b). This makes it extremely difficult, if not impossible, to test them empirically (Cummins 1991a). Unequivocal research results to support or refute these hypotheses are not available in the Netherlands, or anywhere else (cf. Baker & De Kanter 1983; Cummins 1991a; Driessen & van der Grinten 1994; Lucassen & Köbben 1992; Willig 1985). Generally, the interdependency hypothesis suffers from a one-sided focus on language skills while the general cognitive development and the socio-economic and pedagogical family climate are ignored (de Jong, Mol & Oirbans 1988).
- Apart from this general criticism, the interdependency hypotheses cannot be proven right or wrong with regard to at least half of the non-indigenous children in the Netherlands because the (official native) language they have to learn during MTI is not their mother tongue, but a foreign (third, or even fourth) language. For all Moroccan children for example, the MTI language is Standard-Arabic. This is a formal, largely written language which bears little or no resemblance to their actual home language, in this case one of the many Moroccan-Arabic dialects or Berber varieties (Otten & de Ruiter 1993). A similar situation applies to the Kurds, who in the Netherlands receive instruction in Turkish, which is the language of their oppressors in Turkey, and also to other population groups who speak a minority language or a dialect in their country of origin.

- As time passes, minority children increasingly speak Dutch both intra- and intergenerationally (Driessen 2004). This irrevocably means that the proportion of pupils to which the above-mentioned 'educational psychology–argument' potentially applies, is rapidly diminishing.
- Specific conditions may apply with reference to Standard-Arabic. Educationalists point out that learning this language during the phase of alphabetization can interfere with the learning of Dutch. This is due to the fact that Arabic is written in the opposite direction (from right to left), uses different symbols, and does not register certain vowels (Schippers & Versteegh 1987).

Educational theory arguments

For

- It is often assumed that education should build as closely as possible on the knowledge and skills the pupil has already acquired. At school, a continuation takes place of the socialization process of the child, which up until that time largely took place within the family. In view of the fact that minority children speak the language of their country of origin at home and infants are still largely dominant in this language, education – at least partly and during the starting phase – should make room for the home language. The use of the home language will not only allow the learning process to run more smoothly, but will also help to bridge the gap between home situation and the (Dutch) school. Because of the fact that the school is also paying attention to the 'own' culture, many minority parents will feel that they are being more highly appreciated and will get the impression that the school is respecting them. In this respect MTI fulfils a psychological function towards the minority parents.
- Within this concept, it is the most important task of the MTI teacher to instruct pupils in their own language and culture, but in addition, he or she can also fulfill an important role as an intermediary between the school and the parents. With the teacher as an intermediary, the minority parents can no doubt more easily get involved in Dutch education.

Against

- It is held against this line of argument that the future of the minority children lies in the Netherlands. Even though their parents sometimes still dream about migrating back to their country of origin, it appears that the children themselves want to stay in the Netherlands. It is therefore of little use to burden them unnecessarily with extra educational content. The latter is all the more true, because precisely these children are already lagging way behind in Dutch education (Driessen 2001). If the children attend MTI during school hours, they subsequently miss parts of the regular Dutch curriculum. The lessons they have missed are only partly compensated for. There is evidence that the pupils who do not make up for these lessons do significantly less well in Dutch and arithmetic than

the pupils who do catch up on them. If the children attend MTI after school hours, often in combination with many hours of Koran instruction, then this, according to their teachers, can lead to them being so tired during Dutch education that they are no longer able to fully concentrate (Driessen 1994a).

- For Moroccan children (and in general for all children whose home language deviates from the official native language) the argument according to which MTI is necessary to bridge the gap between school and home environment does not apply. As Standard Arabic is in fact a foreign language for everyone, and given that a large percentage of the parents have little or no command of this formal, largely written language, instruction in standard Arabic might even widen the gap instead of narrowing it. Furthermore, those who do have a command of Standard Arabic largely belong to a different cultural group than the illiterate. And in view of the fact that it is the official Moroccan culture that is being passed on there is also very little room, if any, for the culture of the largest group of Moroccans in the Netherlands, i.e. the Berbers. The situation is similar for the Kurds and other cultural minority groups.
- The task of the MTI teacher is often fairly complicated, because he or she has to try and steer a middle course between different groups, each having their own demands and expectations with respect to MTI (Jungbluth & Driessen 1989). In addition, it is difficult for teachers to assume a mediating role because their command of the Dutch language often leaves a lot to be desired (Driessen, Jungbluth & Louvenberg 1988; Inspectie van het Onderwijs 1988).

Social psychology arguments

For

- Within the context of education it is necessary to pay attention to the home language in order to ensure a favorable development of the personality, a positive self-image, and to give children a sense of security and well-being. For many minority groups the home language forms one of the key values of their perception of identity. Insufficient attention to the self-image during childhood can give rise to psychological problems in later life.

Against

- There is hardly any scientific evidence supporting the assumption that MTI has the supposed psychological function. Extensive research into the self-image of children shows that non-indigenous children do not have a particularly negative image of themselves, rather, the opposite appears to be the case (cf. de Jong 1987; Verkuyten 1988; Verkuyten & de Jong 1987). There is therefore no reason for giving separate education to enhance the self-image (de Jong, Mol & Oirbans 1988).
- The concept of culture and identity underlying the pro-argument is too static. It wrongly implies that a child can only choose between one or the other culture and completely

ignores the possibility of an integration of both without a resulting split personality. And what culture should be affirmed? The culture of the parents which they in the past experienced in the country of origin, the 'official' culture in the country of origin, the dynamic mixed culture which the children are currently experiencing in the Netherlands, or an assumed culture of the future (Lucassen & Köbben 1992)? Taking all of this into consideration, it remains to be seen how MTI could possibly contribute to the development of a positive self-image.

Communication arguments

For

- Proficiency in the home language is of major importance in order to follow developments in the country of origin and to gain access to the cultural heritage of the country of origin. In addition, a good command of the home language is important for both intra- and intergenerational communication in the Netherlands as well as for communication with the country of origin.

Against

- With regard to getting access to written sources of culture in the country of origin via MTI, the main criticism is that the expectations are generally too high. The level which can be achieved via MTI is not very high (see below). For a large part of the minority children the level is probably too low to enable them to read magazines or books in the home language with any degree of ease. Besides, it is questionable whether sources really have to be read in the home language as clearly more and more such publications are being translated into Dutch. In addition, individuals can also get acquainted with their 'own' culture in Dutch, via modern means of communication such as television, video, audio equipment and computers.
- It is also argued that children should be able to communicate with their family members in the country of origin by means of writing. Here the fact is often ignored that an important part of these family members – both in the country of origin and in the Netherlands – are illiterate (Driessen 1991b; 1993) and therefore not able to read or write letters. This applies to Moroccans and Turks in particular. And again, not many children achieve a level of linguistic proficiency which enables them to read or write letters without major effort.
- As in the socio-psychological argument the fact is often overlooked that both culture and language are not static but dynamic concepts. As a result of years of stay in the Netherlands the non-indigenous languages and cultures will change (cf. de Jong, Mol & Oirbans 1988). The vocabulary, for example, is influenced by the Dutch (and also English) language. Thus the gap between the 'own' language and culture in the country of origin and in the Netherlands will probably widen.

Language-policy arguments

For

- Multi-lingualism should not be seen as a temporary problem, but as a permanent source of enrichment of a society. In that sense it has an intrinsic function. Multi-lingualism is a plus not only from a social and cultural, but also from an economic point of view. Trade and industry can greatly benefit from multi-lingualism, in particular when it comes to international contacts.
- Furthermore, by supporting MTI the government shows that it attaches importance to the languages and cultures of minorities. This can enhance their sense of self-esteem.
- Finally, via MTI the government also meets national and international guidelines on language rights.

Against

- Languages can probably be learned in a more efficient and effective way than via MTI. Evaluations, after all, show that there is hardly any direct link between MTI-participation and language proficiency. Furthermore, the number of minorities who will actually take up jobs in the international economy is only a fraction of the number of children receiving MTI. From this point of view MTI can therefore be regarded as a bad investment.
- The Dutch government can enhance the sense of self-esteem equally well, or perhaps even better – should the need for this exist – by providing regular education that has been well-adapted to the needs of minority pupils.
- As far as the (inter)national guidelines are concerned, it should be emphasized that these are only guidelines and not obligations do not give individuals a legal entitlement to instruction in a minority language (Extra, Folmer & van der Heijden 1992; Driessen 1994a).

Emancipation arguments

For

- The Dutch language can best be learned via the home language. In this way the minority child will be able to get the most out of education, and, furthermore, the disadvantage at the start of formal education is eliminated as quickly as possible. MTI in this way fulfils a crucial role in achieving equal opportunities. Via MTI the government also expresses the principle of equality of cultures, which is essential for the emancipation of minorities in Dutch society. Sometimes it is argued that MTI offers compensation for the current form of education which tries to re-educate non-indigenous children by means of subtle adaptation strategies, i.e. to adapt them to middle-class values. MTI is therefore essential to combat assimilation.

Against

- At least for the Netherlands it has never been shown that MTI can contribute to increasing educational opportunities. In view of the fact that the educational psychology arguments (see above) cannot be applied to at least half of the pupils, there is no reason to suppose that MTI also works in a facilitating fashion. Considering the actual low home language proficiency level of many of the pupils, it is unlikely that it positively influences the learning of Dutch (cf. Driessen 1994b).
- One important point of criticism in relation to MTI is that strengthening a separate, 'ethnic' identity can hamper the necessary integration of the minority children into Dutch society. It leads to segregation and is therefore anti-emancipatory (Lucassen & Köbben 1992).

5. Implementation and Practical Problems

Altogether there was plenty of discord and doubt about aims and functions of MTI in the Netherlands. Furthermore, evaluation research showed that there were major problems in relation to the practical realization of MTI. Three main problem areas were identified:

Availability and training of teachers

There was not only a major shortage of teachers, but their qualifications were also found wanting. Although the majority of teachers had undergone teacher training, this had usually taken place in the country of origin. Teaching methods sometimes strongly deviated from those generally used in the Netherlands. Thus, for example, whole-class instruction with an emphasis on imitating, learning by heart, and obedience was practiced instead of differentiated instruction with an emphasis on individual development and independent work. The Dutch-language-proficiency level of the MTI teachers was a major problem. According to the majority of school principals it was often so low that communication between teachers was extremely limited.

Limited resources and underdeveloped methods

As a result of the vague and frequently changing objectives of MTI formulated by the Dutch government, it was extremely difficult for educational practitioners to determine exactly what should be pursued. This insecurity was reflected in curricula and teaching methods. MTI-teachers were in fact relatively free to decide how they wanted to shape MTI; i.e. to interpret aims and directions, content and depth of the subject matter. Although a fair amount of material had come onto the market, teaching methods and educational resources were still underdeveloped. There was a particular shortage of educational resources suitable for the Dutch situation. The MTI methods did not connect with those used in regular education, neither with regard to contents, nor with regard to didactics. Furthermore, these reflected the multi-cultural reality in the Netherlands only sporadically, and the educational quality was well below the mark. Despite the fact that the composition

of the MTI classes was very heterogeneous, hardly any possibilities for differentiated teaching existed. Finally, considering the basic infrastructure, only half of the MTI-teachers had their own classroom; the rest of them had to sit somewhere in a corner in the hall, in the teachers' room or some other unsuitable area. And even if classrooms were available, these were mostly insufficiently equipped.

Insufficient integration with regular education

A major problem was that MTI was not embedded in and integrated with regular education. First of all there was discord about the importance of MTI and about the interpretation of its aims. Each of the groups directly involved – minority parents, MTI teachers, the Dutch teaching staff – held their own, often opposing ideas, which could put MTI teachers in an awkward situation. In addition, every now and then, those not directly involved – the media, public opinion, politicians, and scientists – also became involved in the discussion.

All in all, MTI from the very start did not stand a chance within regular education. The position of MTI was further worsened by language barriers, by deviating teaching strategies, by insufficient resources, and by the fact that MTI-teachers frequently worked at more than one school, which made it difficult for them to take part in team-meetings and consultations. For an important part, its standing was also determined by the lack of interest or even negative attitude of the rest of the school team, which saw MTI primarily, and often exclusively, as the responsibility of the MTI-teachers, and not of the school.

A growing unfavourable political climate

Of course, the position of MTI cannot be separated from the situation of minorities in Dutch society at large. Under the influence of the recent economic recession resentment about perceived inequality, unfair treatment and discrimination has taken hold. At the same time, it is becoming clear that under present conditions the Dutch welfare state is no longer affordable. This has led to increasing pressure on individuals in the most unfavorable positions. With regard to the minorities, the government now places priority on a quick acquisition of the Dutch language (and culture). This is seen as the key factor for participation in the labor market. Under given circumstances there is less and less room for 'extra' provisions such as MTI. As the educational position of Turks and Moroccans in particular is very unfavorable and does not seem to have improved very much over the years (cf. Driessen 1992a, 1992b, 1993, 2000, 2002), pressure is growing to improve their integration.

6. Evaluation Research on the Effects of Mother Tongue Instruction

Despite the fact that bilingual education (reception programs and MTI) was offered for 35 years, only very few evaluation studies were carried out in this field. Furthermore, there is a fair amount of discussion about whether the research methodology applied in the existing studies was adequate. In this respect, the situation in the Netherlands does not differ a great deal from that in other countries (e.g., Crawford 1997; Lam 1992).

Bilingual education was evaluated with regard to two possible effects: first, effects on the command of the mother tongue and knowledge of the native culture, and, second, effects on the proficiency level in the Dutch language and other aspects of the regular curriculum. The review at hand hopes to provide better insights into these effects with a focus on the proficiency level in the mother tongue. The reason for this is that in most programs the goal of retaining the mother tongue carried more weight than the role of the mother tongue as a support language. Studies were selected according to the following criteria: (1) The focus was on Turks and Moroccans; there is hardly any material on other language groups. (2) Studies pertain to children in primary education; this type of education covers a period of eight years in total, the first two are nursery education and the last six years are actual 'school years'. The children concerned are usually aged between four and twelve. (3) Finally, the studies investigate a correlation between bilingual education and home-language proficiency.

Study 1: Teunissen (1986)

Model: The study evaluated an experimental bilingual model in which, in the third grade, Turkish and Moroccan children were taught in their home language during 55 per cent of the time and in Dutch during 45 per cent of the time. In the fourth grade, the ratios were reversed. Both years were spent in ethnically homogeneous classes. From the fifth grade onwards, the children attended regular Dutch education (together with Dutch children) and additionally received 2.5 hours of MTI a week. A control group received only regular Dutch education and 2.5 hours of MTI a week.

Aim: The study investigated the effects of a bilingual-bicultural educational program for Moroccan and Turkish pupils.

Design: Longitudinal, quasi-experimental, five cohorts.

Sample: A local sample was drawn comprising a bilingual group with 70 Turkish pupils at one primary school and another group with 78 Moroccan pupils at another primary school. Both schools were located in the same city. The control group comprised 86 Turkish pupils from 20 primary schools. Children were about six to seven years old at the start of the study. When drawing samples, matches were made on starting level, age, participation in nursery education, social background, sex, and home language.

Period: 1980-1984.

Instruments: Two receptive oral tests on listening comprehension and vocabulary were used to match the bilingual group with the control group. Three tests on vocabulary, word decoding and reading comprehension were used to measure effects. These were parallel tests, i.e. Turkish or Arabic translations of existing Dutch tests or self-developed tests. Reliability computations were not made for all of the tests; the number of items involved in each of the tests is not known either. Attention was paid to concurrent validity. By means of the latter three tests, the level of comprehension, verbal fluency and reading proficiency

was tested in Turkish and Arabic respectively; writing proficiency was not tested. The tests for the matching were administered prior to the experiment (at the end of the second grade), the tests for the effect measurement itself in the third and fourth grade (i.e. during the experiment) and in the two subsequent grades, i.e. the fifth and sixth grade. Not all of the tests, however, were administered to all of the pupils in all of the grades. One Arabic test (reading comprehension) could not be administered at all because it turned out to be too difficult. On average, 30 to 40 pupils from each sub-group took part in the tests. Pupils repeating a year were tested a year later and their results were subsequently included in those of their original cohort.

Techniques: Table analysis, analysis of variance.

Results: At the beginning of the experiment, vocabulary and listening comprehension of the Turkish and Moroccan bilingual groups were not different from those of the Turkish control group. In the vocabulary, word decoding and reading comprehension, tests the Turkish bilingual group, in all grades, scored significantly higher than the Turkish control group. The Moroccan bilingual group scored significantly lower than the Turkish bilingual group in practically all grades and on all tests.

As far as the development of language proficiency is concerned, the vocabulary test for the Turkish bilingual group showed an average change in score from 14.7 to 15.7 points (SD 0.2) and the Turkish control group from 12.3 to 12.5 (SD 0.3); the Moroccan bilingual group moved from 11.9 to 14.5 points (SD 0.3). In the decoding test, scores for the Turkish bilingual group changed from 19.4 to 62.7 (SD 0.1); for the Turkish control group from 16.4 to 50.2 (SD 0.1) and for the Moroccan bilingual group from 19.8 to 24.0 (SD 0.6). It should be kept in mind that these development data pertain to a period of four years.

Teunissen concludes that the bilingual program had a clear and fairly constant positive effect on the development of the native language of minority pupils, without seriously affecting their development in the Dutch language and related school subjects.

Comments: For the Turkish bilingual model there was a control group; it was not possible, however, to find such a comparative group for the Moroccan bilingual model. Consequently, it is hard to interpret the test results of the Moroccan children. Teunissen compares the scores on the Turkish tests with those on the Arabic tests. However, it is debatable whether this comparison is reasonable. Furthermore, it is not clear exactly which pupils were tested at what moments and with the aid of which instruments. The interpretation which can be yielded from the test results is that the language development of the Moroccan pupils lags behind that of both the Turkish bilingual group and the Turkish control group. The positive effects noted by Teunissen therefore only apply to Turkish bilingual classes.

Study 2: Verhoeven (1987)

Model: The study evaluated two experimental bilingual models with Turkish pupils in the third grade of school. In model A children were first taught in Turkish but only for a period of two months. After this, they attended lessons given in Turkish in separate mother tongue classes for about half of the time, and the other half of the time they attended lessons given in Dutch together with Dutch pupils. By the end of the fourth grade, Turkish and Dutch children all received full-time education in Dutch; the Turkish children, however, had some extra hours of MTI.

In school B pupils were only taught in Turkish; instruction in Dutch was not added before the fourth grade. School A practiced a variant of the simultaneous model, and school B a transitional model. A control group of Turkish children in third and fourth grade also attended a few hours of MTI a week in addition to Dutch education.

Aim: The study investigated the effects of two models on the acquisition of language proficiency in Turkish and Dutch.

Design: Longitudinal, quasi-experimental.

Sample: Local samples in four cities were drawn. The bilingual group attended two primary schools in grades 3 and 4 (17 and 8 Turkish pupils, respectively). A control group was formed of children attending ten primary schools in grades 3 and 4 (a total of 74 Turkish pupils). The average age at the start was 6.5. When drawing the sample, matches were made on various factors such as home language, social background, participation in nursery education, age, and the number of grades that had to be repeated.

Period: 1982-1984.

Instruments: Four tests were administered to establish oral proficiency: phoneme discrimination (30 items), receptive vocabulary (108 items), productive vocabulary (80 items), sentence imitation (24 items). Three tests measured reading proficiency: word recognition (reading out words correctly), word spelling (32 items), reading comprehension (20 items). Parallel versions were developed for all of these tests: one in Dutch and one in Turkish. The tests were used to measure oral proficiency, reading and comprehension skills; writing proficiency was not included. The reliability (KR20, α) of the tests varied from 0.83 to 0.96. Attention was paid to concurrent validity and content validity. The tests were administered three times: after one month, after ten months and after twenty months. However, as a result of a number of pupils dropping out of the experiment (because of remigration and having to repeat a year), only seventy per cent of the original group were tested at the last measurement.

Techniques: T-test, multivariate analysis of variance, regression analysis, correlation analysis.

Results: At measuring moment 1, the oral proficiency levels of pupils in Turkish and Dutch were ascertained. Both the bilingual and the control group appeared to be signifi-

cantly better in Turkish than in Dutch. As far as the development of oral Turkish proficiency during the three measuring moments is concerned, it became apparent that both groups showed significant progress at about equal rates; however, the bilingual group already started with higher scores at measuring moment 1.

As far as reading proficiency is concerned, there were no significant differences between the experimental group and the control group in word recognition and word spelling; the pupils from the bilingual group, however, were better at Turkish and the pupils from the control group better at Dutch. As regards reading comprehension, there were no significant differences between the two groups in Turkish or Dutch.

Verhoeven's final conclusion was that the education received by the pupils in the control group did not adequately measure up with the background of the Turkish pupils, in view of the fact that for these pupils Turkish is still the dominant language - even after two years of nursery education in Dutch. When comparing the two groups, it became apparent that the Turkish pupils in the bilingual model were achieving better results in Turkish and more or less comparable results in Dutch.

Comments: Parallel versions were developed for each of the tests used; in general, this meant that the Dutch tests were 'translated' into Turkish. Next, the results obtained in both tests were compared. This approach is based on the assumption that the underlying psycho-linguistic operations for both languages are identical, or at least comparable; it is questionable whether this assumption is in fact justified.

Due to pupils having to repeat a year as well as due to migration back to the homeland, both groups showed an experimental drop-out rate of about thirty per cent after one year. It is not clear how this affected the average level of the remaining pupils.

One remarkable conclusion (which was not drawn in the research report) is that, taking into account the differences in scores at the start of the research, it apparently makes little or no difference for the level of Turkish whether pupils (in grade 3) are taught in Turkish for half of the time (the transitional model A), all of the time (simultaneous model B) or for merely a few hours a week (the control group). In other words: the amount of instruction does not appear to affect the level achieved.

The research does not answer a number of important questions. Thus the effect of MTI received by the pupils in the control group remains unclear, and the relevance of class sizes is not discussed: If the two bilingual classes were (much) smaller than the control group classes – as the report implies – then it is possible that effects are partly due to the fact that the children in the first group received much more individual attention from their teachers than the ones in the second group.

Study 3: Driessen, de Bot & Jungbluth (1989)

Model: MTI – the overwhelming majority of the children received MTI, but the number of hours per week varied greatly.

Aim: The study examined the correlation between MTI participation on the one hand and knowledge of the home language and culture as well as performance in 'regular' Dutch education on the other hand.

Design: Cross-sectional, quasi-experimental.

Sample: National random sample of 120 primary schools comprising all pupils in the eighth grade; 368 Turkish and 254 Moroccan pupils; average age: 12.5.

Period: 1987/88.

Instruments: Written language tests (multiple choice and open questions) in Turkish (81 items) and Standard Arabic (53 items) were used. As far as the linguistic description level was concerned, distinctions were made in the test between pragmatics, idiom, vocabulary, grammar and spelling. The level of the tests was adjusted to the abilities expected of the pupils according to those involved in MTI (MTI teachers, policy makers, educational and language experts). Pupil self-evaluation scales were used with scores ranging from (1) 'I cannot do this' to (5) 'I find this very easy' (12 items). Teachers rated pupils ranging from (1) 'definitely no command' to (5) 'definite command' (4 items). The reliability (α) of the instruments ranged from 0.83 to 0.93. Face validity, concurrent validity and content validity were studied in more detail. The tests were used to measure reading and writing proficiency. The pupil self-evaluation scales and teachers' ratings were used to measure listening comprehension, verbal fluency, reading and writing proficiency. Written language tests were used in Dutch language and mathematics (35 and 25 multiple choice items; reliability 0.87 and 0.86).

Techniques: Table analysis, correlation analysis, analysis of variance, regression and path analysis.

Results: Turkish pupils got 73 per cent of the Turkish test items right. Of the Moroccan pupils, 42 per cent did not get a single item of the Arabic test right; for the remaining 58 per cent, the average percentage of correctly answered items was 33. For Turkish and Moroccan pupils, the scores on the self-evaluation scale were 4.2 and 3.2 respectively. It should be noted that for the Moroccans the verbal command was better than the written command (3.5 and 3.0, respectively). On the teachers' rating scale the Turkish and Moroccan pupils scored 4.1 and 3.5, respectively.

In order to study the correlation between MTI participation and home language proficiency, MTI participation was indexed using the average number of hours per week during which the children over the past few years had received MTI. Because of this, it was necessary to check first whether there were any differences between the various categories of MTI participation. The correlation between MTI participation and language proficiency was linear in only a few cases. For the Turkish pupils, the maximum nominal-metric correlation (eta) with the three effect measures (test, scale and rating) was 0.27 and for the Moroccans it was 0.26.

The following characteristics appeared to be of (some) importance in explaining the differences in home language proficiency. For the Turkish pupils they were MTI participation, age (negative), motivation, the importance parents attach to MTI, the MTI teacher working towards the aim of 'language maintenance'. For the Moroccan pupils the most important factors were language use at home (Moroccan Arabic versus Berber), the length of stay (negative), use of Dutch (negative), and the degree of contact between the MTI teacher and the pupils' parents. Speaking a Berber dialect at home in particular had a strong negative effect on the command of Arabic. The linear correlation (*r*) between home language proficiency and Dutch language proficiency for Turkish pupils amounted to a maximum of 0.14 and, for Moroccan pupils, -0.22 (negative). As to the correlation between MTI participation and Dutch language and math proficiency, a distinction was made between MTI during school hours and MTI after school hours. The correlation for the Turkish group was not significant while for the Moroccan group it was negative (-0.11 and -0.13). For the children who attended MTI after school hours the correlation was negative (-0.25 and -0.18 for the Turks, and -0.15 and -0.17 for the Moroccans).

The final conclusion drawn by Driessen et al. is that there are major differences in language proficiency levels. The level of Turkish appears to be fairly good, whereas the level of spoken Arabic is poor and that of written Arabic is definitely very low. The correlation between home language level and MTI attendance is weak to very weak. This also applies to the correlation with Dutch language and math proficiency.

Comments: During the construction phase, the Standard Arabic test had to be simplified significantly because a pilot study had shown that the pupils would not be able to attain the level originally aimed for; in addition, the test was produced in vocalized Arabic, which also meant a simplification.

It can be seen as a disadvantage that one of the instruments used – the language test – only measured general language proficiency and was not sub-divided into sub-tests; this makes it hard to establish the level of the pupils on certain psycho-linguistic sub-aspects.

In a comparative study in three Moroccan cities involving 117 pupils (grades 2-4; aged between 8-12) the Standard Arabic test was administered again (Bentahila & Davies 1990). 47, 51 and 78 per cent of the pupils in the second, third and fourth grade were able to answer the test questions correctly. Bentahila & Davies concluded that the results in the Netherlands were not all that bad compared with the results obtained in Morocco itself. According to Driessen & de Bot (1991), the findings in Morocco shed a new light on the low test results obtained in the Netherlands. On the one hand, the level being aimed at in the Netherlands is far too high, on the other hand, the level achieved in Morocco itself is very low.

Study 4: van de Wetering (1990)

Model: MTI. All children received MTI; the extent to which they received MTI varied widely.

Aim: The study investigated the effectiveness of MTI for Moroccan children, especially the effects of education in the Arabic language.

Design: Longitudinal, quasi-experimental.

Sample: Local sample, eight primary schools in two large cities; 447 Moroccan pupils from grades 3-8, aged between 6-14.

Period: 1983-1985.

Instruments: Decoding test (63 words), reading comprehension test I (14 multiple choice items), reading comprehension test II (13 multiple choice items). The decoding test measured reading proficiency and verbal fluency, and the reading comprehension tests measured the level of reading proficiency. Pupils were not required to produce any written work. All three tests were translated and edited versions of existing Dutch tests. The level of the tests was adjusted to the level at the end of third grade of primary school. In this study, the reliability of the tests was not established, but reference was made to studies in which the reliability (α) for reading comprehension was 0.63. Aspects of concurrent validity were taken up. The tests were administered three times, however, not always on the whole sample since the experiment had a drop-out rate of over forty-five per cent.

Techniques: Table analysis.

Results: The researcher presents the test results in relation to the number of years of Arabic-language instruction the pupils had received both in the Netherlands and in Morocco. In the word decoding test 71 per cent of the pupils with 3 or more years of Arabic education got at least 33 of the items right in the first year of the study. In the second year, the same score was achieved by 72 per cent of the pupils with at least 4 years of Arabic education, and in the third research year by 76 per cent of the pupils with at least 5 years of Arabic education. In reading comprehension test I, 70 per cent of the pupils with 3 or more years of Arabic education got at least 10 out of 14 questions right in the first year; in the second year the same score was achieved by 84 per cent of the pupils with 4 or more years of Arabic education, and in the third year by 87 per cent of the pupils with at least 5 years of Arabic education. In reading comprehension test II, 39 per cent of the pupils got 10 or more of the 13 questions right in the first year; in the following year the same score was achieved by 33 per cent of the pupils, and in the last research year by 48 per cent of the pupils.

Van de Wetering's final conclusion was that the yield of MTI in the field of word decoding and reading comprehension is relatively low, and that 3 years at least are needed to round off the decoding process. Only after 5 to 6 years of MTI can pupils be expected read out and comprehend a simple vocalized text, on the condition, however, that MTI is allowed to function under reasonably favorable circumstances.

Comments: It should be noted that no more than 54 per cent of the pupils participated in the decoding test; for reading comprehension test I, this percentage was 24 and for reading comprehension test II it was a mere 9 per cent. This was partly caused by the fact that MTI

teachers often did not allow their pupils to participate in the test because they felt their pupils could not meet the required level. Additionally, pupils who were not successful in the first or second test were not allowed to go on with the second or third test. In addition, about half of the respondents dropped out during the course of the study. Furthermore, the number of test items was extremely small and only reading proficiency was tested; no research was carried out about writing proficiency. Consequently, the results as presented in the study undoubtedly overestimate the level of the group as a whole. The interpretation of van de Wetering is not correct or at least incomplete. From re-analyses it appears that the level of language proficiency is extremely low and that this level deteriorates for each consecutive grade as the years go by.

Study 5: Aarssen, de Ruiter, & Verhoeven (1992)

Model: MTI. The pupils participated in MTI to various degrees.

Aim: The study´s objective was to develop tests which could be used to assess the proficiency of individual pupils in Turkish and Arabic at the end of primary school. On the one hand, the results of these tests could be used to evaluate what they had learned during MTI, on the other hand the results provided useful information for the mother-tongue teachers at secondary schools.

Design: Cross-sectional, quasi-experimental.

Sample: The study consisted of three parts: an exploratory study, the main study (both in the Netherlands) and a comparative study (in Turkey and Morocco). The exploratory study pertains to a local sample including 23 primary schools in seven cities with 69 Turkish and 81 Moroccan pupils in eighth grade. For the main study, a national random sample was drawn from 68 primary schools and with 263 Turkish and 222 Moroccan pupils in eighth grade. The comparative study in Turkey pertains to a local sample survey at six primary schools in two cities with 276 pupils in the highest grade. The study in Morocco pertains to a local sample from six primary schools in two cities and with 242 pupils in the highest grade.

Period: 1990, 1991.

Instruments: Two tests to establish listening proficiency were applied: vocabulary and instruction comprehension (tests of whether pupils understood teachers' instructions, e.g. 'Mark the left square'). Four tests aimed to establish reading proficiency: word decoding, spelling, vocabulary, syntax and reading comprehension. In the first instance, a total of almost 300 multiple-choice items were used to determine receptive language proficiency, i.e. only comprehension and reading proficiency. In the exploratory study, two parallel versions were developed, a Turkish test and a Moroccan Arabic/Standard Arabic test. The reliability (α) of the original tests for Arabic ranged from 0.38 to 0.93 and the reliability for Turkish tests from 0.57 to 0.92. After adjustment (by means of removing about 100 items),

the reliability varied from 0.67 to 0.94. Part of the tests had been derived from Dutch material intended for six year olds. Content validity and concurrent validity were ascertained.

Techniques: T-test and correlation analysis.

Results: Test scores were converted into the percentage of items answered correctly. In the exploratory study and the main study carried out in the Netherlands, the scores of the Turkish pupils varied between 52 and 96, with scores in listening proficiency between 75 and 87, and scores in reading proficiency between 52 and 67. The word-decoding test was not included in the latter; its scores ranged from 93 to 96. In the comparative study carried out in Turkey, the scores were generally higher than those in the main study (5 to 18 test points), with the exception of reading comprehension and instruction (6 points lower). The researchers concluded that in the Netherlands pupils did reasonably well in the spelling test and fairly satisfactorily in the reading comprehension test. On the other parts, they achieved very reasonable to good scores. Furthermore, the researchers argued that the children in the Netherlands did not lag very far behind those in Turkey, the areas in which the pupils lagged behind most were spelling and written vocabulary.

The Moroccan children's scores in the exploratory and the main study varied between 24 and 81. For listening proficiency, scores were between 52 and 74, and for reading proficiency between 24 and 60. Here again, scores for word decoding were considerably higher, ranging from 81 to 79. In the comparative study in Morocco, the scores on all parts were considerably higher, i.e. 13 to 42 points compared to the main study. The conclusions were that listening proficiency was fair and that instruction comprehension was of a lower level. Few difficulties were encountered in word decoding, and the researchers also considered the level of reading comprehension to be satisfactory. However, the children did not yet have a command of word spelling and syntax and their written vocabulary was not very extensive. According to the researchers, it was possible to conclude on the basis of the differences between the results obtained in the Netherlands and those obtained in Morocco that the children in the Netherlands have a better command of the language than earlier research had suggested (in this case Driessen et al. and van de Wetering).

The final conclusion of this study was that Turkish pupils in the Netherlands have in general a high level of listening proficiency and reading proficiency in Turkish; the listening proficiency of the Moroccan children is reasonably good, but their reading proficiency is fairly poor.

In the exploratory study, the correlation between language proficiency and MTI participation was also examined. For Turkish pupils (who had attended MTI for six years on average), the correlations (r) were between -0.01 and 0.31, whereby the correlations for the oral tests were the highest. The relationships between the degree to which the children spoke Turkish at home and with friends (versus Dutch) were all weak to very weak, and sometimes even negative; the correlations varied from -0.18 to 0.28. Tests of the differences between first grade entrants and higher grade entrants (side streamers) showed that the group that had attended Dutch education from the first grade onwards in general scored

worse. For the Moroccan children the correlations between the test parts and MTI participation varied from -0.13 and 0.30. The correlation between the degree to which the pupils spoke Arabic at home or with friends (as opposed to Dutch) and the test parts was also weak; results varied from -0.16 to 0.38. In general, the differences in language proficiency between first grade entrants and higher grade entrants were small and not significant at any point, which led the researchers to the conclusion that the education received in Morocco (in contrast to educational experience in Turkey) apparently does not have an effect on the test scores.

Comments: The tests used were limited to receptive language proficiency; the pupils were not expected to actively practice language. Therefore, the researchers only succeeded in finding out something about the listening and reading skills (the easier skills), but nothing about the verbal and writing skills (the more difficult skills). In view of this limitation, it can be argued that the researchers went too far by concluding that the Turkish children had a 'good' command of Turkish and the Moroccan children a 'reasonably good' command of verbal Arabic and a 'fairly poor' command of written Arabic. If the children had been subjected to productive tests, their level would undoubtedly have turned out to be considerably lower than suggested by the researchers.

Another point of criticism relates to the fact that in three out of seven Moroccan tests the pupils did not or only just score above guessing level (four multiple choice items with on average 24 to 33 per cent of the items answered correctly). With regard to the results of the exploratory study the picture was too positive. As the research report indicates, only half of the pupils took part in some parts of the test. The remaining pupils were not allowed to participate in the tests by their MTI teachers, who felt that the tests were far too difficult and that their pupils would not be able to do them. If the teachers were right we have to assume that, if these pupils had undergone the tests, the average would have dropped considerably.

Study 6: Wagenaar (1993)

Model: The study examined the effects of an experimental bilingual model in an ethnically homogeneous class in first and second grade of primary education (nursery school) where Moroccan Arabic was spoken in the mornings and Dutch in the afternoons (for 15 and 8 hours per week respectively). Originally the intention was to reverse this in third grade, but owing to the non-availability of Moroccan teachers, it became necessary to switch, after the beginning of second grade, to Dutch education with some additional weekly hours of MTI. The third grade did, however, remain ethnically homogeneous. In fourth grade the children were divided up into various ethnically heterogeneous classes. A control group consisted of Moroccan pupils receiving normal Dutch education as well as some hours of MTI a week.

Aim: The study examined the effects of bilingual nursery education on the school careers of Moroccan children.

Design: Longitudinal, quasi-experimental.

Sample: A local sample was drawn from one primary school (bilingual model) and two primary schools (control group) all in one large city. Two times (in the bilingual and the monolingual school respectively) 30 Moroccan children from the first, second and fourth grade; two times 23 pupils who spoke Moroccan Arabic at home, and two times 7 pupils who spoke Berber. The average age at the start of the study was 4.5. In the sampling, matches were made on ethnic background, age, sex, and socio-economic background.

Period: 1987-1991.

Instruments: The sub-tests for language comprehension and language production (speaking) of the Reynell Developmental Language Test and the sub-tests for passive and active vocabulary of the Language Test for Non-indigenous Children were used. These tests were used to measure oral language proficiency; they were Moroccan Arabic translations of existing Dutch tests. No information is available about the content and number of items of the tests; and no data is presented about the reliability and validity of the tests. The tests were administered three times: at the beginning of the first grade (pre-test), at the end of the second grade (effect measurement), and at the end of the fourth grade (post-test).

Techniques: Analysis of (co)variance, regression analysis, correlation analysis, discriminant analysis.

Results: The results of the first two measuring moments are expressed in age equivalents (The test had been standardized with each test score being replaced by an equivalent age score in years and months.). In the first measurement, the bilingual group scored 3.8 years for Moroccan Arabic language comprehension and the control group scored 4.1 years (the norm is 4.5 years). There was only one significant effect of participating in the experiment with regard to the children's mother tongue: the scores of the Arabic-speaking children were 3.8 and 4.6 years, those of the Berber-speaking children were considerably lower, i.e. 2.5 and 2.4 years in the bilingual group and the control group respectively. The language production level for Moroccan Arabic was much lower, for the bilingual and control group the scores were 2.8 and 2.9 years, which is more than 1.5 years below their norm. Here too, the ethnic/linguistic group had a significant effect. The Arabic children obtained scores of 2.9 and 3.2 years and the Berber-speaking children obtained scores of 2.2 and 1.9 years. In all cases, the scores for Dutch language comprehension and language production were about 2.5 years, which is below the level of Moroccan Arabic, and 2 years below the norm. According to Wagenaar, the latter provides an argument for offering bilingual education.

In the second measurement, which took place when the children were aged 6, the bilingual group's score for Moroccan Arabic language comprehension was 5.3 years and the control group's was 3.7 years. After controlling for the first measurement, this difference turned out to be a significant one. There was also a significant difference between Moroccan-Arabic-speaking and Berber-speaking children: 5.7 years and 4.5 years versus 4.9 and 2.9 years for the bilingual group and the control group, respectively. Therefore, the bilingual group was more than six months and the control group two years below the norm. In the

bilingual group the Berber children were more than one year below the norm. The picture for language production was a comparable one although the overall level of verbal fluency was much lower than the level of comprehension. The scores for the bilingual group and the control group were 3.9 and 2.9 years. After controlling for the first measurement, this difference appeared to be a significant one. There were also major differences between Moroccan-Arabic-speaking and Berber-speaking children, but the differences were statistically not significant. In the bilingual group both sub-groups appeared to have profited from the education received, but the Arabic-speaking children considerably more so than the Berber-speaking ones. Wagenaar states that the proficiency level of the Berber children is still so low that one should have serious doubts about the use of Moroccan Arabic education for this group. She attributes this result to the fact that the home language of the Berber children is a Berber variant and not Moroccan Arabic.

Regression analysis was used to establish which factors could explain the differences in the level of Arabic. For language comprehension ethnic origin (Moroccan Arabic versus Berber) appeared to be the most important predictive factor, followed by participation or non-participation in the experiment (ß=-0.48, respectively ß=0.28). The same factors applied for language production, although the strength of the coefficients differed (ß=-0.26 and ß=0.61, respectively). IQ, sex, educational level of parents and language orientation at home no longer appeared to have a significant effect.

In the third measurement, which took place two years after the experiment was completed, the expected norm for passive vocabulary was between 83 and 92. Although the scores of the bilingual group were considerably higher than those of the control group (62 versus 47), this difference turned out to be not significant. Both groups stayed well below the norm, despite the fact that both groups had attended MTI in the third and fourth grade. The expected norm for active vocabulary was between 42 and 51; the scores of the bilingual group and control group were 18 and 17 respectively. After controlling for the starting measurement, this difference was not significant either. Wagenaar concluded that the home-language proficiency of the bilingual group continued to develop well during the period in which the experiment was run, but stagnated when the experiment was stopped. In particular, productive language proficiency appears to be sensitive to education. According to the researcher, the Moroccan Arabic language is given insufficient support in the home environment for a continued development.

As to proficiency in Dutch, Wagenaar concludes that its development was not hindered by bilingual education, although no clear positive effects were found. This conclusion is not on par with the finding that a relatively large number of children from the bilingual group were referred to special education. Discriminant analysis showed that this was mainly related to the Berber background and the very traditional values within the family on the one hand, and to participation in the bilingual model, i.e. attending Moroccan Arabic education, on the other. On this basis, Wagenaar advises Berber-speaking children (which are by far the majority of the children of Moroccan origin in the Netherlands) against participating in this type of bilingual nursery school.

Comments: It is unfortunate that there are no norms available for the tests used, which were all Moroccan Arabic translations. This makes it virtually impossible to put the results in the right perspective.

7. Summary and Conclusions

It is remarkable that in 35 years of bilingual education no more than six studies were financed in which the effect of this type of education on home-language proficiency was examined. In fact, only one study (by Driessen et al. 1989) of the effects of MTI was financed by the Ministry of Education. Probably this was due to the controversial and sensitive nature of this type of education. Like in any matters of minority policies, politicians and policy makers were, until recently, reluctant to take definitive stands on the issue (cf. Lucassen & Köbben 1992). It is nevertheless remarkable that no need was perceived for gaining insights into the effects of a considerable investment on the part of the state (in financial terms) and the children concerned (more than 10 per cent of the time spent in primary education).

All the evaluation studies reported here have methodological shortcomings. Ideally, studies should have a longitudinal design, they should adequately assess characteristics at the starting point of the experiment, study sufficiently large groups of pupils, verify all (educational) activities undertaken within the framework of the intervention and apply tests which adequately differentiate between various language modalities and psycho-linguistic sub-skills. As has been shown above, there are several methodological flaws in the studies discussed. In this respect, the Dutch situation is not peculiar (see e.g. Baker & de Kanter 1983; Birman & Ginsberg 1983; Willig 1985). Consequently, it is almost impossible to make well-founded statements on the effectiveness of bilingual programs. Statements about the level of language skills as such are somewhat less risky to make, but not entirely unproblematic either.

Taking this into account and leaving aside instructional characteristics, the following trends in language skills can be observed: The level of – both oral and written – Turkish appears to be reasonably good. As far as the level of Arabic is concerned, an emphatic distinction has to be made between Moroccan Arabic (the informal, spoken language) and Standard Arabic (the formal, written language). The command of the first is limited, of the second downright poor. This can largely be attributed to the fact that to all Moroccans Standard Arabic is a foreign language. Furthermore, the children tested do not live in Morocco, but in the Netherlands, which makes it even more difficult for them to learn Standard Arabic:

The effects of MTI and bilingual programs on the level of language proficiency do not become entirely clear. Transitional and simultaneous models appear to be more effective than MTI, which is probably related to the length of the period of instruction, the age at which it is provided, the length of stay in the Netherlands, and the home language of the children. Regarding long-term effects of transitional and simultaneous models, the results

are not promising (Wagenaar 1993). The correlations between MTI and home-language proficiency are weak at very best, and negative at worst. For the language level, factors such as home language (e.g., Moroccan Arabic versus Berber or Dutch) and length of stay (the longer the period of stay in the Netherlands, the lower the level) appear to be at least as important as the (period of) instruction. Among Moroccan children in particular, a definite loss of language seems to take place, or perhaps more aptly stated: a stagnation in language acquisition. One should probably not expect MTI to achieve more than a slowdown of this process.

In the 1980s and 1990s, MTI was expected to contribute to a higher level of proficiency in Dutch and thus more promising school careers of minority pupils. Furthermore, bilingual programs were first introduced in the Netherlands because it was argued that immigrant pupils had the right to learn their home language (also referred to as 'mother tongue,' 'native language' or 'first language') in order to be able to communicate with their family in that language, gain access to their own cultures, or acquire this language as an aid to facilitate the learning of Dutch. Such expectations have not been fulfilled – certainly not for the Moroccans – and cannot be fulfilled.

One major obstacle to the fulfillment of some of these aims was the absurd insistence on teaching a large group of immigrant children a language that was in fact not the language they spoke in their families. It hardly comes as a surprise that forcing Berber-speaking children to learn Standard Arabic does not help their acquisition of the Dutch language. The main question here is why this obviously illogical strategy was ever introduced and pursued over a considerable period of time.

Political aims were generally vague and changed over time, so that the implementation of an unambiguous program and its application over a longer term were impossible. Policies represented a number of compromises (and were thus partly inconsistent) and lacked a scientific foundation.

The history of the Dutch academic debate about bilingual education can be characterized as a battle between linguists on the one hand, and educational sociologists on the other. To a certain extent, the disagreement between the two scientific disciplines can be traced back to the old controversy about deficit versus difference. Linguists tend to stress the importance of learning languages as a goal in itself and as enrichment for the individual and society. Sociologists tend to emphasize educational opportunities and life chances of minorities. According to some Dutch sociologists, investing in bilingual programs was a complete waste as there was no evidence suggesting that such programs would be successful. As they argued, this was unlikely as most minorities were second or third generation immigrants who increasingly spoke Dutch at home and whose future lay in the Netherlands. With hindsight, it seems that sociologists have defeated linguists in this battle.

References

Aarts, R. (1994). Functionele geletterdheid van Turkse kinderen in Turkije en Nederland. De Lier: ABC.

Aarssen, J., de Ruiter, J., & Verhoeven, L. (1992). Toetsing Turks en Arabisch aan het einde van het basisonderwijs. Tilburg: Tilburg University Press.

Aarssen, J., de Ruiter, J., & Verhoeven, L. (1993). Summative assessment of ethnic group language proficiency. In: G. Extra & L. Verhoeven (Eds.), Immigrant languages in Europe. Clevedon: Multilingual Matters, 159-179.

Bataens Beardsmore, H. (1993). European models of bilingual education: practice, theory and development. Journal of Multilingual and Multicultural Development, 14, 103-120.

Baker, C. (1988). Key issues in bilingualism and bilingual education. Clevedon/Philadelphia: Multilingual Matters.

Baker, C. (1990). The effectiveness of bilingual education. Journal of Multilingual and Multicultural Development, 11, 269-277.

Baker, K., & de Kanter, A. (Eds.) (1983). Bilingual education. A reappraisal of federal policy. Lexington: Lexington Books.

Birman, B., & Ginsberg, A. (1983). Federal policy and the efctiviness of blingual education. In: K. Baker & A. de Kanter (Eds.), Bilingual education. A reappraisal of federal policy. Lexington: Lexington Books, ix-xxi.

Broeder, P., & Extra. G. (1995). Minderheidsgroepen en minderheidstalen. Den Haag: VNG Uitgeverij.

CALO (1992). Ceders in de tuin. Naar een nieuwe opzet van het onderwijsbeleid voor allochtone leerlingen. 's-Gravenhage: DOP.

Crawford, J. (1997) Best evidence: Research foundations of the Bilingual Education Act. Washington, DC: National Clearinghouse for Bilingual Education.

Cummins, J. (1991a). Conversational and academic language proficiency in bilingual contexts. In: J. Hulstijn & J. Matter (Eds.), Reading in two languages. AILA Review 8. Amsterdam: Free University Press, 75-89.

Cummins, J. (1991b). Interdependence of first- and second language proficiency in bilingual children. In: E. Bialystok (Ed.), Language processing in bilingual children. Cambridge: Cambridge University Press, 70-89.

Demirbaş, N. (1990). OETC-beleid; theorie en praktijk. Utrecht: NCB.

de Bot, K., Driessen, G., & Jungbluth, P. (1991). An evaluation of migrant language teaching in the Netherlands. In: K. Jaspaert & S. Kroon (Eds.), Ethnic minority languages in education. Amsterdam/Lisse/Berwyn PA: Swets & Zeitlinger, 25-35.

de Jong, M.J. (1987). Herkomst, kennis en kansen. Lisse: Swets & Zeitlinger.

de Jong, M.J., Mol, A., & Oirbans, P. (1988). Zoveel talen, zoveel zinnen. De behoefte aan lessen Eigen Taal in het V.O. Rotterdam: EUR.

de Ruiter, J.J. (1989). Young Moroccans in the Netherlands: an integral approach to their language situation and acquisition of Dutch. Utrecht: RUU.

Driessen, G. (1990). De onderwijspositie van allochtone leerlingen. De rol van sociaal-economische en etnisch-culturele factoren, met speciale aandacht voor het Onderwijs in Eigen Taal en Cultuur. Nijmegen: ITS.

Driessen, G. (1991a). Landstaal of moedertaal? Het problematische karakter van de 'eigen taal' binnen het Marokkaanse OET(C). Migrantenstudies, 7, (2), 2-14.

Driessen, G. (1991b). Ontwikkelingen in T1- en T2-vaardigheidsniveau van Turkse en Marokkaanse leerlingen. Toegepaste Taalwetenschap in Artikelen 41, (3), 27-35, 76.

Driessen, G. (1992a). Etnische herkomst, verblijfsduur, thuistaal en taalvaardigheid Nederlands. Toegepaste Taalwetenschap in Artikelen 44, (3), 7-19, 111.

Driessen, G. (1992b). First and second language proficiency: prospects for Turkish and Moroccan children in the Netherlands. Language, Culture and Curriculum, 5, 23-40.

Driessen, G. (1993). Socio-economic or ethnic determinants of educational opportunities? Results from the educational priority policy programme in the Netherlands. Studies in Educational Evaluation, 19, 265-280.

Driessen, G. (1994a). Naar een meer realistische benadering van het Onderwijs in Eigen Taal en Cultuur? Pedagogische Studiën, 71, 47-59.

Driessen, G. (1994b). Moroccan children acquiring Arabic in the Netherlands. In: G. Driessen & P. Jungbluth (Eds.), Educational opportunities. Tackling ethnic, class and gender inequality through research. Münster/New York: Waxmann, 71-88.

Driessen, G. (1995). Ontwikkelingen met betrekking tot het Onderwijs in Eigen Taal (en Cultuur): beleid, argumenten en perspectieven. Toegepaste Taalwetenschap in Artikelen, 53, (3), 95-107, 228.

Driessen, G. (1995). The educational progress of immigrant children in the Netherlands. Language, Culture and Curriculum, 8, 265-280.

Driessen, G. (1996). Minority Language and Culture Teaching in the Netherlands: policies, arguments, evaluation and prospects. Compare, 26, 315-332.

Driessen, G. (1997). From mother tongue to foreign language: Prospects for minority-language education in the Netherlands. In: Th. Bongaerts & K. de Bot (Eds.), Perspectives on foreign-language policy. Studies in honour of Theo van Els. Amsterdam/Philadelphia: Benjamins, 181-200.

Driessen, G. (2000). The limits of educational policy and practice? The case of ethnic minority pupils in the Netherlands. Comparative Education, 36, 55-72.

Driessen, G. (2001). Ethnicity, forms of capital, and educational achievement. International Review of Education, 47, 513-538.

Driessen, G. (2002). School composition and achievement in primary education: A large-scale multilevel approach. *Studies in Educational Evaluation*, 28, 347-368.

Driessen, G. (2004). De taalsituatie van Caribische en Mediterrane immigranten. Ontwikkelingen in taalvaardigheid en taalkeuzes in Antilliaanse, Surinaamse, Turkse en Marokkaanse gezinnen gedurende de periode 1995-2003. Migrantenstudies, 20, 74-93.

Driessen, G., & de Bot, K. (1990). Turkse en Nederlandse taalvaardigheid en leerlingkenmerken. Leerderskenmerken: individuele verschillen in het leren van talen. Toegepaste Taalwetenschap in Artikelen 37, (2), 83-91, 137.

Driessen, G., de Bot, K., & Jungbluth, P. (1989). De effectiviteit van het Onderwijs in Eigen Taal en Cultuur. Prestaties van Marokkaanse, Spaanse en Turkse leerlingen. Nijmegen: ITS.

Driessen, G., Hulsen, M., Aarssen, J., & Cohen de Lara, H. (2003). OALT als taalondersteuning in de onderbouw van het basisonderwijs. Voortgangsverslag. Nijmegen/Utrecht: ITS/Sardes.

Driessen, G., Jungbluth, P., & Louvenberg, J. (1988). OETC in het basisonderwijs. Doelopvattingen, leerkrachten, leermiddelen en omvang. 's-Gravenhage: SVO.

Driessen, G., & Withagen, V. (1998). Taalvariatie en onderwijsprestaties van autochtone basisschoolleerlingen. Taal en Tongval, 50, 2-24.

Driessen, G., & van der Grinten, M. (1994). Home language proficiency in the Netherlands. The evaluation of Turkish and Moroccan bilingual programmes - A critical review. Studies in Educational Evaluation, 20, 365-386.

Driessen, G., van der Slik, F., & de Bot, K. (2002). Home language and language proficiency: a large-scale longitudinal study in Dutch primary schools. Journal of Multilingual and Multicultural Development, 23, 175-194.

Extra, G., Folmer, J., & van der Heijden, H. (1992). Tweetalig basisonderwijs: Modellen, argumenten en ervaringen. Tilburg: KUB.

Eldering, L. (1989). Ethnic minority children in Dutch schools: Underachievement and its explanations. In: L. Eldering & J. Kloprogge (Eds.), Different cultures, same school. Ethnic minority children in Europe. Amsterdam: Swets & Zeitlinger, 107-136.

Fase, W. (1987). Voorbij de grenzen van onderwijs in eigen taal en cultuur. Meertaligheid op school in zes landen verkend. 's-Gravenhage: SVO.

Fase, W. (1994). Ethnic divisions in Western European education. Münster/New York: Waxmann.

Inspectie van het Onderwijs (1988). OETC: Niet apart maar samen. 's-Gravenhage: Ministerie van Onderwijs en Wetenschappen.

Jungbluth, P., & Driessen, G. (1989). Onderwijs in Eigen Taal en Cultuur. Pretenties aanzienlijk, verwachtingen gering. Pedagogische Studiën, 66, 52-60.

Lam, T. (1992). Review of practices and problems in the evaluation of bilingual education. Review of Educational Research, 62, 181-203.

Lindholm, K. (1990). Bilingual immersion education. Criteria for program development. In: A. Padilla, H. Fairchild & C. Valadez (Eds.), Bilingual education. Issues and strategies. Newbury Park: Sage, 91-105.

Lucassen, L., & Köbben, A. (1992). Het partiële gelijk. Controverses over het onderwijs in de eigen taal en cultuur en de rol daarbij van beleid en wetenschap (1951-1991)]. Amsterdam/Lisse, Swets & Zeitlinger.

Malakoff, M., & Hakuta, K. (1990). History of language minority education in the United States. In: A. Padilla, H. Fairchild & C. Valadez (Eds.), Bilingual education. Issues and strategies. Newbury Park: Sage, 27-43.

Ministerie van Onderwijs en Wetenschappen (1991). Eigen taal als onderdeel van een geïntegreerd talenonderwijs. 's-Gravenhage: DOP.

Ministerie van Onderwijs, Cultuur en Wetenschappen (1995). Uitwerkingsnotitie Onderwijs in Allochtone Levende Talen. Zoetermeer: MOCW.

Otten, R., & de Ruiter, J. (1993). Moroccan Arabic and Berber. In: G. Extra & L. Verhoeven (Eds.), Community languages in the Netherlands. Amsterdam/Lisse/Berwyn PA: Swets & Zeitlinger, 143-174.

Schippers, A., & Versteegh, K. (1987). Het arabisch. Norm en realiteit. Muiderberg: Coutinho.

Snow, C. (1990). Rationales for native language instruction. Evidence from research. In: A. Padilla, H. Fairchild & C. Valadez (Eds.), Bilingual education. Issues and strategies. Newbury Park: Sage, 60-74.

Teunissen, J. (1986). Een school, twee talen. Een onderzoek naar de effecten van een tweetalig-biculturee onderwijsprogramma voor Marokkaanse en Turkse leerlingen op basisscholen te Enschede. Utrecht: RUU.

Tweede Kamer (2003). Rapportage Integratiebeleid Etnische Minderheden. Tweede Kamer der Staten-Generaal. Vergaderjaar 2003-2004, 29203, nrs. 1-2. 's-Gravenhage: Sdu Uitgevers.

van de Wetering, W. (1990). Onderwijs in Eigen Taal en Cultuur aan Marokkaanse kinderen in Nederland. Het OETC als resultante van een maatschappelijk krachtenspel. Utrecht: RUU.

van der Grinten, M., & Driessen, G. (1993). OETC in het basisonderwijs, een review. Stand van zaken in 1992/1993 op basis van empirische gegevens. Nijmegen: ITS.

van der Slik, F., Driessen, G., & de Bot, K. (2000). Thuistaal en taalvaardigheid in het basisonderwijs: een longitudinaal onderzoek. Gramma/TTT - Tijdschrift voor Taalwetenschap, 8, 119-144.

van het Loo, M, de Spiegeleire, S., Lindstrom, G., Kahan, J., & Vernez, G. (2001). A comparison of American and Dutch immigration and integration experiences. What lessons can be learned? The Hague: WRR.

Verhoeven, L. (1987). Ethnic minority children acquiring literacy. Tilburg: KUB.

Verhoeven, L. (1994). Transfer in bilingual development: The linguistic interdependency hypothesis revisited. Language Learning, 44, 381-415.

Verkuyten, M. (1988). Zelfbeeld van allochtone jongeren niet negatiever dan bij autochtonen. Didaktief, juni, 7-9.

Verkuyten, M., & de Jong, W. (1987). Zelfwaardering en onderwijsleerprestaties van Turkse kinderen. Pedagogische Studiën, 64, 498-507.

Vermeulen, H., & Penninx, R. (Eds.)(2000). Immigrant integration: The Dutch case. Amsterdam: Het Spinhuis.

Wagenaar, E. (1993). Tweetaligheid in het aanvangsonderwijs. Een onderzoek naar de effecten van tweetalig kleuteronderwijs op de schoolloopbaan van Marokkaanse kinderen. Amsterdam: Het Spinhuis.

Willig, A. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. Review of Educational Research, 55, 269-317.

Mother Tongue Teaching and Programs for Bilingual Children in Sweden

Monica Axelsson

Introduction

In Sweden, mother tongue instruction in public schools for children of migrant background has, since the middle of the 1970s, been firmly established. Unquestionably, the education of children with a first language other than Swedish is a relevant issue: Currently (2004) 14 per cent of all children in pre-school and at school have a first language other than Swedish. In the suburbs of the major cities Stockholm, Göteborg and Malmö, the share of language-minority children in schools sometimes amounts to between 50 and 100 per cent. In 2000, almost twenty per cent of Sweden's 8.9 million inhabitants had either themselves been born abroad or had at least one parent born abroad. This situation is largely due to post-1970 immigration. Major countries of origin include Finland, Norway and Denmark, the former Yugoslavia, Iraq and Iran.

While there is a long tradition of mother tongue teaching, it has repeatedly been the subject of controversial debates. Evaluations of its effects on the overall educational performance of the children, however, remain scarce and limited in scope. In particular, longitudinal studies on bilingual children's overall academic achievement at school are lacking. This paper will first provide an overview of the political framework and the quantitative development of mother tongue instruction in Swedish schools. In the second part, major results of evaluations of the effects of such programmes on children's academic achievement will be presented.

The Development of Mother Tongue Education in Sweden

The development of mother tongue education in Sweden has to be seen against the background of public discussions at the beginning of the 1970s and the Immigrant Commission's report (SOU 1974) which led to the introduction of a general immigrant and minority policy in 1975. Here, the Swedish Riksdag decided to base migration and integration policy on the principles of equality, partnership and freedom of choice. The latter principle implies a rejection of cultural assimilation and underlies the right to mother tongue education. In 1977 the "home language reform" followed in response to a public campaign for immigrant rights in Swedish society. It gave immigrant and minority children the legal entitlement to mother tongue instruction in public schools. The reform was based on two di-

verse motives: on the one hand it aimed to provide support for the retention of ethnic or cultural identity, on the other, it was intended to allow for a normal linguistic, academic and cognitive development of immigrant and minority students. This concern for the individual student's cognitive and academic development during school years was the main motive. Altogether, the reform was intended to help ensure equality between Swedes and persons from other ethnic backgrounds as well as cultural freedom of choice.

The Immigrant Commission's report (SOU 1974) reflected ideas already expressed in the UNESCO recommendation from 1953. This recommendation stated that all children should get their initial schooling in their mother tongue (UNESCO 1953: 68ff). It was based on the obvious fact that knowledge cannot be acquired in a language the student does not understand. Modern learning theories also assume that new knowledge can be acquired more easily if instruction is based on the known. Learning should thus start from the situation, the knowledge and proficiencies an individual already has. Based on these assumptions, mother tongue instruction has enjoyed the support of all major political parties in Sweden. Furthermore, it has since the 1970s been widely accepted that:

- it takes long, several years[1], before a second language works as efficiently for learning as the first language,
- an abrupt language shift will hamper a child's cognitive development during school years and diminish his or her potential academic achievement,
- children's attitudes about their first language – and through that also to their own culture and origin – will be affected by the role this language is accorded in society and at school,
- an individual's identity growth relies upon a person's experiencing positive affirmation of his or her cultural background (after Hyltenstam & Tuomela 1996: 31).

Mother tongue instruction expanded along with the general expansion of school education. During the later 1970s, the number of bilingual students increased, and the government provided funds earmarked for mother tongue instruction. This led to a situation in which municipalities and schools benefited from a high number of bilingual students and subsequently more bilingual teachers[2] were employed (Municio 1996). During the 1980s, the number of bilingual students decreased, but the government continued to subsidise L1 instruction. With falling numbers of bilingual students and a round of budget cuts, however, the schools had to choose between employing regular Swedish teachers or bilingual staff and they usually opted for the former. Many bilingual teachers became unemployed, and L1-instruction no longer expanded at the previous pace.

---

[1] More recent research has confirmed and specified that it commonly takes about five to eight years for a second language learner to arrive at a level of a monolingual student (Cummins 1981; Collier 1987; Thomas & Collier 1997, 2002; Hakuta, Butler & Witt 2000).

[2] By this term I refer to teachers specifically employed to give mother tongue instruction to these bilingual students.

In 1990, an evaluation of mother tongue instruction by Riksrevisionen (the State audit institution) included criticism of both the costs involved in L1 instruction as well as organizational aspects. The report received a lot of publicity, and before it reached Parliament, many municipalities in Sweden cut down on mother tongue instruction, making a large part of L1-teachers redundant. In the end, Parliament did not accept the recommendations of the report but amended the home-language reform: The provision of mother tongue teaching was no longer obligatory if fewer than five students in the municipality requested it.[3]

Moreover, in 1991, the responsibility for public schools passed from the Swedish government to the municipalities. New principles for financing schools were introduced, and teaching time, like all other matter of expenses for education, had to be meticulously calculated. Earmarked funds for L1 instruction no longer existed, rather, schools or headmasters were free to assign resources according to their own priorities. It was no longer economically beneficial for schools to offer instruction for language-minority students.

As a result of these developments, instruction in immigrant languages dramatically lost in importance. While in 1990 96,000 teaching-hours per year were spent for L1 instruction, in 1994, the figure was down to 39,500 teaching-hours per year. The decrease in L1 support has also affected pre-school education: While in 1980 64 per cent (11,072) of children with another L1 than Swedish got L1 support, in 2000 this was true for only 13 per cent (5,044). Municio (1996) argues that the disappearance of earmarked funds and the possibility for the schools to decide about priorities were the main causes for the reduction in L1 instruction. Additionally, organisational problems and the prevailing disagreement about methods of instruction played a role. However, new regulations have also made it easier to establish schools with bilingual instruction. Thus by the year 2000, nine free or independent Swedish/Finnish schools and 15 Islamic schools with bilingual instruction in Arabic and Swedish had been founded.

At present, children and youths who in their daily lives communicate with at least one parent in a language other than Swedish are, under certain conditions, entitled to mother tongue education (National agency for education 2003). It is the parent, in communication with the child, who generally requests L1-instruction from the school. The subject "Swedish as a second language (SSL)" has, since 1995, had a curriculum parallel to "Swedish". The qualifications acquired in SSL can form part of the exams necessary for entering higher education including universities. According to the National curriculum, the headmaster has a special responsibility to ensure that "the instruction and caring for students is organised in such a way that students get the specific support and help they need" (Lpo 94: 19).

---

[3] Since 1994 the national minorities of Saami, Tornedalians and Rom are exempted from this rule as well as from the precondition that the language has to be used on a daily basis at home in order to get instruction at school. In June 1999 the newly enacted state policy on national minorities gave Meänkieli, Saami, Yiddish, Romani and Finnish recognition as national minority languages.

For the past five years, 2000-2004, the Swedish government has provided special funds for the major cities Stockholm, Göteborg and Malmö in order to support their development and combat the segregation of immigrants in dense urban areas. In all three cities, this has led to the emergence of a variety of efforts to enhance bilingual children's linguistic and academic achievement. In Stockholm in-service training has been offered to all categories of teachers in pre-school and school in form of the course "Bilingual children's language acquisition and academic achievement". The course is at the university level and covers the themes migration, culture and communication, bilingual development, performance assessment and academic achievement in a second language. In evaluations of this course', teachers stated that they:

- developed new approaches on how to work with bilingual children aimed at promoting language ability and successful academic development
- now show more respect for the children's bilingualism and multicultural competence
- now allot more time for interaction and production
- have increased collaboration with other teachers and parents
- use performance assessment (Sellgren 2002).

Moreover, the heads of these pre-school and school teachers notice a change in teacher behaviour after the course. The heads report to have observed:

- a change in teachers' attitudes towards children and parents
- more encouragement of L1 and increased collaboration with parents
- more dialogue and interaction with the children
- performance assessment and with it a new way of looking at Swedish as a second language
- a strengthened focus on language in subjects like math and science
- new approaches to and new choices of texts
- a new organisation of the teaching of Swedish and Swedish as a Second Language
- increased confidence and respect from colleagues
- positive effects on children and parents (Sellgren 2002).

Evaluations of this and other urban projects have been carried out by researchers from various disciplines. In Stockholm, an evaluation of language development and school achievement (Axelsson et al. 2002) has resulted in a rewriting of both pre-school and school plans. Both plans stress that "Facilitating the development of the language ability of children/students with Swedish as a second language must mean to actively support their command of their mother tongue as well as to develop the second language, Swedish" (Stockholm pre-school plan 2003: 9 and Stockholm school plan 2004: 8). As the school plan further emphasizes, knowledge about multilingual development throughout the school

system needs to be extended in order to improve the process of second language learning. Furthermore, the school plan assumes that bilingual students' academic achievement depends upon the development of successful mother tongue instruction courses including teaching of content matter in the mother tongue (Stockholm school plan 2004:8).

A recent report from the National Agency for Education (Skolverket) "More Languages, More Opportunities" (2002) gives an overview of the development and current state of mother tongue instruction:[4]

- Ten years ago, 60% of multilingual children received mother tongue assistance at preschool. In 2002, the respective figure was 13%.
- At compulsory comprehensive schools, the percentage of students entitled to and receiving mother tongue instruction has fallen from 60% to 50%.
- Most mother tongue instruction is given in the afternoon – at the close of the regular school day.
- Only 10% of Sweden's municipalities arrange instruction in minority languages in particular subjects.

In contrast to critical views and in opposition to financial cuts in the 1990s, the report calls for an increase in mother tongue instruction. One important reason for this is the unsatisfying situation of language-minority students in compulsory schooling: more than every fifth student with a mother tongue other than Swedish in 2002 failed to qualify for a national programme at upper secondary school. In comparison, the same is true for only ten per cent of all students (The National Agency for Education 2003).[5] This raises several questions: Did Sweden receive more immigrant children since 1999? Did the bilingual children not receive any instruction in Swedish as a second language? Did the bilingual children receive any mother tongue instruction? Did the bilingual children receive any adapted content instruction? And if they did, what quality did this instruction have? Unfortunately, these questions cannot be answered for the individual child. Furthermore, it is difficult to identify the individual influence of each factor on the educational achievement of bilingual children. Even a study with experimental design and with matched control groups could hardly grasp the complexity of the institutional and individual conditions. The study by Hill (1995) described below attempts to answer some of the questions listed above based on qualitative research.

[4] www.skolverket.se, "Tema Modersmål". In 1997 the terms 'home-language' and 'home-language instruction' were changed to 'mother tongue' and 'mother tongue instruction'.

[5] For several decades school results for students with a foreign background have been lower than for majority students. Statistics show that Swedish majority students have a more successful schooling than "students with a foreign background". This lable is used for children born outside Sweden or children born in Sweden, but with both parents born outside Sweden.

Studies on the Effects of Mother Tongue Instruction

Margret Hill (1995): Opportunities for immigrant children. On academic and language development in pre-school and school.

Early quantitative studies on the effects of mother tongue instruction (for example Löfgren 1985, 1991) are limited in scope. Here I want to report on a qualitative study from 1995 by Margret Hill of Gothenburg University's Department of Education, that formulates important research questions.

The purpose of the study was to investigate the connection between participation in L1 instruction in pre-school and school and:

- academic performance in year 9 in Swedish, mathematics, English and overall
- well-being at school (at present)
- identity
- attitudes toward the mother tongue and bilingualism
- problems with Swedish (according to the youths' own opinions)
- correctness, fluency and accent in Swedish (according to the interviewer's assessment)
- ability to reason about the abstract concept of democracy.

Informants were 42 17-year-old immigrant adolescents, born in 1977, and interviewed during their first year in senior high school. All of them had attended full or part time pre-school where they had received L1 instruction. That is, they had all participated in L1 instruction from the beginning of pre-school, and they had all had the opportunity to get L1 instruction throughout all school years.

Hill distinguishes three groups according to different patterns of mother tongue instruction:

Group 1: continuing (45%), L1 instruction from pre-school to grade 9

Group 2: late discontinuation (29%), L1 instruction at least until and including grade 5

Group 3: early discontinuation (26%), discontinued L1 instruction before grade 5

The youths lived in suburbs of Göteborg, the second largest city in Sweden, in areas inhabited predominantly by immigrants (Biskopsgården-Länsmansgården, Frölunda-Tynnered, Angered-Bergsjön). Half of them were girls and half were boys. There was an even distribution of theoretical and practical high school programs and an even selection from three senior high schools. The students in the study belonged to ten national groups speaking the following languages: Finnish, Yugoslavian languages, Turkish (including speakers of Assyrian and Kurdish), Spanish (with speakers from Spain and Latin America), Persian, Polish, Arabic (including speakers from North Africa, Lebanon, Iraq and Syria), Portuguese (with speakers from Portugal, Brazil and Cape Verde), Chinese languages and Greek. It should be

noticed that 81 per cent of the students came from the lowest socio-economic group in society.

Results:

Grades: For all 42 students, grades were on the same level as they were for average Swedish students. Their grades in English, however, were slightly higher.

Group 1: all grades were 1-3 tenths higher than for the average Swedish student.

Group 2: grades were close to group 1, but they had slightly lower grades in Swedish. However, their grades were not lower than those of the average Swedish student. In mathematics their grades were a couple of tenths higher.

Group 3: consistently lower grades, especially in Swedish (8 tenths lower than group 2 and a whole grade lower than group 1).

Well-being: The question Hill asked was "How happy are you with the program you have chosen for senior high school?"

Group 1: all students were very happy with the chosen program

Group 2: 58% were very happy, 17% were planning to change the program

Group 3: 55% were "happy/rather content", 45% were considering changing the program or dropping out.

Ethnic identification: The question Hill asked was if they "mainly identified with their own ethnic group" or if they "mainly felt like Swedes".

Group 1: 63% mainly identified with their own ethnic group, 37% mainly felt like Swedes.

Group 2: 50% mainly identified with their own ethnic group, 50% mainly felt like Swedes.

Group 3: 27% mainly identified with their own ethnic group, 73% mainly felt like Swedes.

Attitudes to L1 and bilingualism:

Group 1: 100% held positive attitudes to both

Group 2: 17% held positive attitudes to L1, 50% were positive to bilingualism as such

Group 3: 18% held positive attitudes to L1, a minority held positive attitudes to bilingualism.

Students' own assessment of language proficiency in Swedish:

Group 1: 100% stated that they never or seldom had problems with their Swedish at school.

Group 2: 67% stated that they never or seldom experienced problems with their Swedish.

Group 3: 46% stated that they never or seldom experienced problems with their Swedish.

Assessment of linguistic correctness, fluency and accent in Swedish (interviewer assessment):

Group 1: "Almost all" students spoke fluent, accent-free, correct Swedish.

Group 2: Altogether similar to group 1, but "apparently" lower fluency.

Group 3: Only 45% of the students had flawless Swedish, only 18% spoke without accent and only 27% showed fluency.

Ability to reason around the abstract concept of democracy: During the interview, Hill asked each student to reason around the concept "democracy". The students' speech was analysed according to a model in which their reasoning was categorised into four groups: fragments, loose facts, concepts and generalisations. The results are presented as follows:

Table 1: The students' levels of sophistication in talking about democracy

Way of treatment /group	Group 1	Group 2	Group 3
Fragments	3	1	4
Loose facts	1	2	3
Concepts	5	2	3
Generalisations	9	7	-

Table 1 shows that the more qualified the treatment of the concept "democracy" was, through use of concepts and generalisations, the more often the student belonged to Group 1 or 2 and vice versa. The more fragmented and vague the treatment of the concept "democracy," the more likely the student was to belong to Group 3.

To sum up, students who had received the least L1 instruction during their school years (Group 3) showed the greatest difficulties in Swedish – both according to the interviewer and to their own judgement. Students with the lowest proficiency in Swedish were, in their own judgement and the interviewer's opinion, the ones who felt most like Swedes. There is a clear pattern in Hill's study showing that the most successful students were those who participated in L1 instruction from pre-school through primary school (Group 1). The least successful were those who only experienced limited L1 instruction (Group 3).

However, is it possible to conclude that the first group, i.e. students with continuing L1 instruction all through school, performed well in school because they participated in L1 instruction? The answer has to be no. Other circumstances in the students' and their families' life situations might have been important for the strategies they developed for making choices at school and for their general school success. Possibly students who attended L1 instruction had parents who not only supported their participation in that but were generally more interested in the children's schooling and more supportive. This interpretation is supported by comments from students in Group 3 ("early discontinuations") who, when asked why they discontinued their L1 instruction, often mentioned "sceptical adults". Parents in this group were uncertain about the best instruction model for their children and had possibly interpreted the children's poor language development in Swedish as a result of L1 instruction. Possibly, these parents were influenced by a prevailing discourse in society

that emphasized the priority of learning Swedish. The result for some of the children in Group 3 is that they moved between various instruction models during different stages, sometimes with and sometimes without L1 instruction. Furthermore, it might be the case that the students who participated in L1 instruction throughout their school career had greater talent for and interest in language. They received good grades in Swedish and they had a high proficiency in Swedish according to their own and the interviewer's assessment. It should be noted that they also had good grades in English. Finally, we must acknowledge the possibility that, due to the small sample, results are not representative and biased.

Still, the results of Hill's study are well in line with predominant theoretical assumptions about bilingual development, and they are compatible with other studies such as those by Thomas and Collier (1997; 2002, see also Cummins 2000 for comments on the Thomas & Collier study).

Veli Tuomela (2001): Bilingual development during school years. A comparison of Sweden-Finnish[6] students in three instruction programs

In order to identify possible differences with regard to language development in different instruction programmes in Sweden, Veli Tuomela carried out a quasi-longitudinal study with 180 bilingual students. The students participated in one of three instruction programmes:

Group 1: regular Swedish class with one or two hours of mother tongue support per week,

Group 2: bilingual class,

Group 3: independent Swedish-Finnish school with bilingual instruction.

The aim of the study was to examine the relative degree to which the students in each of the three programmes displayed command of Swedish and Finnish, respectively, with use of the language grammatically correct and with demonstrated competence of the language through use of complex words, varying sentence structure, etc. as a measurement.

Informants were selected according to the following criteria: a) both parents' first language was Finnish, b) the student was born in Sweden or moved to Sweden before he or she reached the age of three, c) Finnish was a language of daily communication in the family. Monolingual students in Sweden and Finland formed control groups. Students attended grades 3, 6 and 9 during the study.

All students were tested in both languages for vocabulary and grammar by tests developed by the researcher. The students further answered a questionnaire on language use in various domains and concerning self-assessment of their proficiency in the two languages.

[6] Sweden-Finnish is the term for the Finnish language variety spoken in Sweden and is coined as a parallel to Finland-Swedish designating the Swedish language variety spoken in Finland.

Results:

Students in the regular Swedish class with mother tongue support (group 1) assessed their Swedish to be more proficient and their Finnish as less proficient than students from the other two instruction programmes. These students also, to a greater extent than the other students, used Swedish in communication with parents, siblings and friends.

The linguistic analyses – error analyses and performance analyses (i.e. what structures the students actually used) – showed that all students could make themselves understood and had a fairly high degree of language proficiency in both languages. The average lexical and grammatical error frequency in both Finnish and Swedish varies between a fraction of a percent and a few percent in most of the age groups across all three programs. There were small differences in linguistic complexity. For all students, linguistic complexity increased over the years. On a quantitative basis, group 1 had somewhat lower measures in Finnish, especially in writing, while the other two groups had lower measures in Swedish. As for linguistic correctness, there were few differences between the groups and error frequency was overall very low in grammar and vocabulary. In spoken Finnish group 1, students made twice as many errors as the other students. As for spoken Swedish, the situation was reversed; in group 2 and 3, students made twice as many errors. In the control groups, students made fewer errors overall in their respective mother tongue than any of the bilingual students. The greatest differences between the groups appeared in written production. Group 1 had more lexical and grammatical errors than the other groups, both in Swedish and Finnish. A somewhat surprising result was shown for spelling in Swedish, as, bilingual students made fewer errors than monolingual Swedish students.

Summary

- No major differences were found between students in bilingual classes (group 2) or independent schools (group 3) in either Swedish or Finnish.
- Similarly, there were no major differences in Swedish between bilingual students in Swedish classes (group 1) and monolingual Swedish students.
- Students in Swedish classes performed better in spoken and written Swedish than students in bilingual classes.
- Despite the fact that Finnish was their mother tongue and was used in the family, students in Swedish schools were better in Swedish than in Finnish.

According to Tuomela, the results correspond to a great extent, with previous studies of the language of Sweden-Finnish students. The results present students with a near-native bilingual competence.

Conclusion and Outlook

Knowledge is primarily mediated through language. Language is a vital ingredient in cognitive processing. The more proficient one is in a language, the more likely he or she is to perform well academically. Furthermore, one's language and mother tongue, in particular, has great importance for the individual's sense of security, confidence and identity. For all these reasons, mother tongue instruction has, since the 1970s, been accorded a central place in the Swedish educational system.

Over the years, mother tongue instruction has been more or less integrated into the regular school schedule. There have been bilingual classes with both languages used side by side or, more common, mother tongue instruction as a subject within or outside of the regular school day. Since 2000 two major trends can be identified: On the one hand, some immigrant groups have begun to establish their own independent bilingual schools, and on the other hand, municipality schools have been giving about 40 minutes of mother tongue instruction per week, outside of the regular schedule. In addition to this, some schools in areas with a majority of bilingual students give some subjects, for example mathematics, in some mother tongues.

Among teachers in Stockholm who have participated in courses on bilingual children's language acquisition and academic achievement, there is awareness about the potentially advantageous resources which a bilingual child possesses such as fluency in the mother tongue and familiarity with more than one culture. These teachers try to develop new approaches on how to work with bilingual children, approaches which better facilitate the development of more than one language, and which better foster academic success in bilingual students. This involves giving more respect to the children's bilingual and multicultural competences and allocating more time for interaction and production in Swedish as well as in the mother tongue. An important part of the development of new ways of educating bilingual children is the establishment of increased collaboration with other teachers and with parents. Non-native speakers are commonly disadvantaged when tested with tests standardised for monolingual students (Cummins 1996), these teachers opt for performance assessment of the student's proficiency in Swedish.

Research on bilingual instruction in public schools in Sweden has so far focused more on language development rather than academic achievement. Hill's study (1995) tries to address the overall situation by investigating bilingual language development, academic achievement, and the students' confidence and identity. It is obvious that bilingual children and their families are sensitive to signals from the surrounding society. These signals indicate recognition of their languages and cultures. Pre-school and school education mirror society, but each measure of the organization, competence and attitudes of the personnel can make a difference. Children benefit from a socio-culturally supportive environment in which each child's competence and experience is the point of departure and in which there is an ongoing dialogue between student and teacher. To insist on only using Swedish in the classroom is a way of cementing already existing power relations. To show respect for

other languages, knowledge and experiences by giving non-native Swedish students space in the classroom is a way of expressing a willingness to care about basic equality in society. This is what Thomas and Collier (1997, 2002) aim to express when they speak about the importance of a socio-culturally supportive context. Existing research supports the assumption that, instead of practising an all-or-nothing ideology for or against mother tongue education, various organizational programmes for instruction in two languages prove feasible. Based on research and practical experience around the world, a list of possible models for implementing instruction programmes focusing on the bilingual students' resources can be drawn up. A school could, in one or several of the languages represented in the school:

- offer bilingual instruction, 50 per cent in the second language and 50 per cent in the mother tongue (cf. Thomas & Collier 1997, 2002),
- teach one or several subjects per school year in the mother tongue (cf. May 1994),
- teach one topic per semester in the mother tongue (cf. May 1994),
- develop co-operation between subject teachers, mother tongue teachers and second language teacher (cf. Bergman et al 2001),
- assign tasks to be solved in the mother tongue and/or in the second language (cf. DeFazio 1997).

The first instruction programme is modelled on the two-way and one-way developmental bilingual education programmes suggested by Thomas and Collier (1997, 2002). In these model programs demands for cognition, learning and social integration with majority children are met. A class in an American two-way developmental bilingual education (BE) program might comprise one half of students with English as mother tongue and the other half of students with Spanish as mother tongue. Instruction in this class is given in English and Spanish in equal shares of time, which means that all students get the opportunity to develop knowledge in their mother tongue as well as in a second language.

In the Swedish school system, a program which is based on assigning tasks to be solved in either L1 or L2 would ideally require that mother tongue teachers are hired locally by the school and that mother tongue instruction is integrated into the regular schedule at favourable hours. All teachers should be involved in a dialogue and cooperative planning in order to give the best possible support to the students. The student should be given the opportunity to fulfil tasks in both of their languages and these tasks should be given an equal assessment and value (cf. Cook 1999). Both classroom and library in the school should be equipped with books and printed material in Swedish and the students' L1 in order to give the bilingual students support to develop academic knowledge in both languages. Such arrangements would also make it apparent to monolingual Swedish-speaking students that knowledge can be mediated through all languages, not only through Swedish.

References

Ada, Alma Flor (1988), The Pajaro Valley experience I: Skutnabb-Kangas, Tove & Cummins, Jim (red.), Minority Education. Clevedon: Multilingual Matters Ltd, 223-238

Axelsson, Monica (1998), En språkpedagogisk utvärdering av arbetet i två internationella klasser i Botkyrka. I: Axelsson, Monica & Norrbacka Landsberg, Riikka, En studie av två internationella klasser ur ett etnologiskt och ett språkpedagogiskt perspektiv. Botkyrka: Mångkulturellt centrum, 61-150.

Axelsson, Monica (1999a), Skolframgång och minoritetsstatus. Skolan – en kraft att räkna med. I: Axelsson, Monica (red.), Tvåspråkiga barn och skolframgång – mångfalden som resurs. Stockholm: Rinkeby Språkforskningsinstitut, 8-35.

Axelsson, Monica (1999b), Skolframgång utan språktillgång – är det möjligt? Språk och makt i den mångkulturella skolan. I: Boström-Andersson, Rut (red.), Ordets makt och tankens frihet. Om språket som maktfaktor. Publikation av föredragen vid Uppsala Humanistdagar den 20-21 mars 1999. Uppsala: Institutionen för nordiska språk, 67-76.

Axelsson, Monica (2000a), Framgång för alla. Från att inte kunna – till att inte kunna låta bli att läsa. I: Åhl, Hans (red.), Svenskan i tiden – verklighet och visioner. Stockholm: Nationellt centrum, HLS Förlag, 9-23.

Axelsson, Monica (2000b), Språkarv, språkidentifikation och språkfärdighet. Alternativ till begreppen modersmål och infödd språknivå. I: Börestam Uhlmann, Ulla (red.), Postskriptum Elsie Wijk-Andersson. Uppsala: Hallgren & Fallgren, 11-28.

Axelsson, Monica (2001), Organisation och lärande i skolor med språklig och kulturell mångfald. I: Axelsson, M., Gröning, I., & Hagberg-Persson, B., Organisation, lärande och elevsamarbete i skolor med språklig och kulturell mångfald. Uppsala: Uppsala universitet, 11-52.

Axelsson, Monica; Lennartson-Hokkanen, Ingrid & Sellgren, Mariana (2002), Den röda tråden. Utvärdering av Stockholms stads stortadssatsning – målområde språkutveckling och skolresultat. Stockholm: Språkforskningsinstitutet i Rinkeby.

Bergman, Pirkko; Sjöqvist, Lena; Bülow, Kerstin & Ljung, Birgitta (2001), Två flugor i en smäll. Att lära på sitt andraspråk. Stockholm: Almqvist & Wiksell.

Collier, Virginia (1987), Age and rate of acquisition of second language for academic purposes. TESOL Quarterly 21, 617-641.

Collier, Virginia & Wayne Thomas (1999a-c), Making Schools Effective for English Language Learners, Part 1-3. TESOL Matters Vol. 9, No. 4-6.

Cook, Vivian (1999), Going Beyond the Native Speaker in Language Teaching. TESOL Quarterly Vol. 33, No.2, 185-210.

Corson, David (1998), Changing Education for Diversity. Buckingham: Open University Press.

Cummins, Jim (1981), Age on arrival and immigrant second language learning in Canada: A reassessment. Applied Linguistics 1, 132-149.

Cummins, Jim (1986), Empowering minority students: A framework for intervention. Harvard Educational Review 56 (1), 18-36.

Cummins, Jim (1994), Knowledge, power, and identity in teaching English as a second language. I: Genesee, Fred (red.), Educating Second Language Children. Cambridge: Cambridge University Press, 33-58.

Cummins, Jim (1996), Negociating Identities: Education for Empowerment in a Diverse Society. Ontario: California Association for Bilingual Education.

Cummins, Jim (2000), Language, Power and Pedagogy. Bilingual Children in the Crossfire. Clevedon: Multilingual Matters.

DeFazio, Anthony (1997), Language awareness at The International High School. In: L. Van Lier & D. Corson (eds.), Knowledge about Language. Vol. 6. Encyclopedia of Language and Education. Dordrecht: Kluwer Academic Publishers, Inc, 99-107.

Faltis, Christian (1997), Joinfostering: Adapting teaching for the multilingual classroom. New Jersey: Prentice Hall.

Gibbons, Pauline (1991), Learning to Learn in a Second Language. USA: Heinemann.

Hajer, Maaike (2000), Creating a Language-Promoting Classroom: Content-Area Teachers at Work. In: Hall, Joan Kelly & Verplaetse, Lorrie Stoops (eds.), Second and Foreign Language learning Through Classroom Interaction. Mahwah: Lawrence Erlbaum, 265-286.

Hakuta, Kenji; Butler, Yuko Goto & Witt, Daria (2000), How long does it take English learners to attain proficiency? The University of California Linguistic Minority Research Institute. Policy report 2000-1. Stanford University. Downloaded from: www.stanford.edu/~hakuta/.

Hill, Margret (1995), Invandrarbarns möjligheter. Om kunskapsutveckling och språkutveckling i förskola och skola. Göteborg: Göteborgs universitet, Institutionen för pedagogik.

Hyltenstam, Kenneth & Tuomela, Veli (1996), Hemspråksundervisningen. I: Hyltenstam, K. (red.), Tvåspråkighet med förhinder? Invandrar- och minoritetsundervisning i Sverige. Lund: Studentlitteratur, 9-109.

Lpo 94 (1998), Läroplan för det obligatoriska skolväsendet, förskoleklassen och fritidshemmet. Skolverket & Fritzes AB.

Lucas, Tamara & Anne Katz (1994), Reframing the Debate: The Roles of Native Languages in English-Only Programs for Language Minority Students. TESOL Quarterly Vol. 28, No. 3, 537-562.

Lucas, Tamara, Henze, Rosemary & Donato, Ruben (1990), Promoting the success of Latino language-minority students: An exploratory study of six High Schools. Harvard Educational Review 60 (3), 315-340.

May, Stephen (1994), Making Multicultural Education Work. Clevedon: Multilingual Matters LTD.

Miramontes, Ofelia (1997), Quality instructional planning for language minority students: A total school commitment. In: Sjögren, Annick (red.), Language and Environment. A Cultural approach to Education for Minority and Migrant students. Botkyrka: Multicultural Centre, 115-126.

Miramontes, Ofelia; Nadeau, Adel & Nancy Commins (1997), Restructuring Schools for Linguistic Diversity. Linking Decision Making to Effective Programs. New York: Teachers College Press.

Municio, Ingegerd (1994), Medpart, motpart eller icke-part? I: Peura, M & Skutnabb-Kangas, T. (eds), Man kan vara tvåländare också. Sverigefinnarnas väg från tystnad till kamp. Stockholm: Sverigefinländarnas arkiv, 18-72.

Ogbu, John (1991), Immigrant and Involuntary Minorities: A Cultural-Ecological Theory of School Performance with Some Implications for Education. I: Gibson, Margaret & Ogbu, John (eds), Minority Status and Schooling. A Comparative Study of Immigrant and Involuntary Minorities. New York: Garland Publishing, 3-33.

Pease-Alvarez, Cindy & Vasquez, Olga (1994), Language socialization in ethnic minority communities. In: Genesee, Fred (ed.), Educating Second Language Children. The whole child, the whole curriculum, the whole community. Cambridge: Cambridge University Press, 82-102.

Skolinspektörernas halvårsrapport 2001. Grundskolan. Stockholms stad: Utbildningsförvaltningen.

Skolplan för Stockholms stad. Ny kurs för Stockholms skolor. Stockholms stad.

Skolverket (2002), Flera språk – fler möjligheter – utveckling av modersmålsstödet och modersmålsundervisningen 2002. [More languages – more opportunities.] www.skolverket.se

Skolverket (2004): Education for students of non-Swedish background and recognized minorities, Stockholm, http://www.skolverket.se/english/system/non-swedish.shtml.

SOU (1974), Invandrarutredningen 3. Invandrarna och minoriteterna. Stockholm:LiberFörlag.

Stockholm pre-school plan (2003), www.stockholm.se/utbildningsforvaltningen

Stockholm school plan (2004), www.stockholm.se/utbildningsforvaltningen

Thomas, Wayne & Collier, Virginia (1997), School effectiveness for language minority students. NCBE Resource Collection Series, No. 9. George Washington University. www.ncbe.gwu.edu/ncbepubs/resource/effectiveness/.

Thomas, Wayne & Collier, Virginia (2002), A National Study of School Effectiveness for Language Minority Students' Long-Term Academic Achievement Final Report. http://crede.ucsc.edu/research/llaa/1.1_final.html

Tuomela, Veli (2001), Tvåspråkig utveckling i skolåldern. En jämförelse av sverigefinska elever i tre undervisningsmodeller. Stockholm: Centrum för tvåspråkighetsforskning, Stockholms universitet.

Verhoeven, Ludo & Vermeer, Ant (1985), Ethnic group differences in children's oral proficiency in Dutch. I: Extra, G. & Vallen, T. (eds.), Ethnic minorities and Dutch as a Second Language. Dordrecht/Holland: Foris Publications, 105-131.

Westin, Charles (2003): Young people of migrant origin in Sweden, in: Emrehan Zeybekoğlu und Bo Johansson (Hg.): Migration and labour in Europe. Views from Turkey and Sweden, Şefik Matbaasi, Istanbul, 170-194.

Wong Fillmore, Lily (1991), When Learning a Second Language Means Losing the First. Early Childhood Research Quarterly 6, 3: 323-346.

Bilingual Development in Primary School Age

Hans H. Reich

Most longitudinal studies of children's bilingual development describe or analyze the individual language developments of children of middle-class background, typically in a situation in which one parent consciously speaks one language to the child and the other parent, the other language, from birth onward (Tracy/Gawlitzek-Maiwald 2000). But this way of growing up bilingually is hardly the only, nor the average situation. The linguistic background of children in immigrant families in Germany presents itself in a quite different light. It is characterized by the usage of the family's native language as the one spoken predominantly in the household and varying usage of the German language spoken outside the family circle. There are some empirical studies of the bilingualism of this population (e.g. Hepsöyler/Liebe-Harkort 1988, 1991; Preibusch 1992; Pfaff 1991; Gogolin/Neumann 1997; Jeuk 2003). But longitudinal research has been scarce up to now.

The following contribution attempts to fill this gap to some degree. It presents the results from a research project about Turkish-German children aged 5-10 years during their primary school years in the city-state of Hamburg. This research was part of a project which aimed at the development of measures for bilingual proficiency levels and, later on, of language support programmes for this group of children (Reich 2004a). It describes very briefly the data collection and the analysis procedure (for more details see Reich 2004b) and then presents some results with regard to the changing relationship between first and second language in grades 1, 2, and 3.

I. Sample and Data Collection

The research involved about 150 children enrolled at 7 primary schools with different percentages of pupils of migrant background. Most of them are born in Germany; all of them have some contact with German, however modest in some cases. Turkish prevails as the language of family communication; in one third of the cases both the parents speak only Turkish to their children. In 5 families Kurdish or Arabic is used alongside Turkish and German with the children.

The children attended different classes, had different teachers and different time tables. Beyond their age and bilingualism, only one common additional feature is to be stressed here, namely, that in all the participating schools at least one Turkish teacher was present so that all children in the sample had some access to Turkish lessons, however widely varying, from class to class and from grade to grade.

One major question for this group was how the language proficiency improved during the time of observation and whether the first and the second language of the pupils were developing in similar or rather different ways. Consequently, when deciding which measurement instruments would be appropriate for our study, the following had to be taken into consideration (1) the necessity to provide for data measuring the children's progress in each language and (2) to allow for a direct comparison between the two languages of the children.

Given the children's age, we opted for a profile analysis of oral texts as the best way to appropriately assess their competencies in both German and Turkish. Around the time of enrolment, around 6 months before school begins, the children were given a picture showing a kitchen scene and a picture story about a cat hunting a bird and the bird's triumphant escape. They were asked to talk about the picture both in German with a German teacher and – at another point of time – in Turkish with a Turkish teacher. The order in which either language was to be used was left free for the teachers to decide. The utterances of both children and teachers were recorded on tape. This procedure was repeated a year later, towards the end of grade 1, with the same instruments and most of the same children.

In order to keep the children motivated to participate, we opted not to repeat the task of explaining the "Kitchen Scene" and the story "Cat and Bird" a second time. In grade 3 they were shown a short video cartoon, "The Mole and his Friends", and asked to tell what they had seen. As this is not the same task as describing the "Kitchen scene" and the picture story "Cat and Bird", comparability in time between results of the latter with the previous tests, i.e. the ability to measure children's progress from grade 1 to grade 3, was reduced to a considerable degree. It remains, however, possible to analyse the relationship of L1 and L2 at this point of their school career and to ask what changes had come about since the end of grade 1.

II. Measures of Bilingualism

For direct comparison of L1 and L2 we decided to use a measure for descriptive and narrative competence. The visual stimuli were divided into elementary parts, it was decided which actors and events were to be mentioned in minimal (understandable) propositions, and university collaborators rated the utterances of the children accordingly. Interrater-reliability was calculated and found to be satisfying.

In addition, as a measure of the size of the child's vocabulary, the verbal types (i.e. different verbs used by the children) were counted. As indicators of grammatical development we analysed the positions of verbs in German and the verbal tense suffixes in Turkish as well as the syntactical means of connecting clauses in both the languages (for details see Reich 2004b). It should be noted that all the measures of vocabulary and grammar are language-specific and, consequently, cannot be directly compared.

III. Before School Start

137 children took fully part in both the Turkish and the German speech tasks at the initial data collection. For each of them a mean score of performance in the kitchen scene and the Cat-and-Bird-story was established for Turkish, and likewise for German. The analysis of these scores reveals that on average the group's narrative competence is clearly higher in Turkish than in German.

Table 1: Language dominance prior to school start

First speech task – means of test scores in Turkish and German at different schools								
	school 1	2	3	4	5	6	7	total
Turkish	2,38	2,41	2,22	2,54	2,32	2,26	2,06	2,32
s	,49	,63	,52	,50	,36	,87	,63	,58
German	1,80	1,23	2,43	1,80	1,81	2,15	2,07	1,88
s	,72	,66	,49	,75	,47	,73	,75	,73

On a scale with a maximum of 4 points the mean score for Turkish is 2.32, and for German 1.88 respectively; Turkish prevails over German by .44 points. We can speak of an unambiguous dominance of Turkish in the group.

Furthermore, the variation in German (SD .73) is considerably wider than in Turkish (SD .58). This reflects the fact that German is acquired neither in similar environments nor according to the children's age but more or less incidentally, dependent on varying input or motivation.

Looking at the performance of the single pupils in both the languages enables us to establish individual language balances. The following table shows the scores for German (categorized) horizontally and the scores for Turkish (categorized in the same way) vertically.

Table 2: Language balances prior to school start

First speech task – categorized levels in Turkish and German: language balances									
level in German	level in Turkish (8 = highest level) 1	2	3	4	5	6	7	8	total
1	1	0	0	0	1	2	0	0	4
2	0	1	2	3	2	5	0	0	13
3	0	0	2	4	13	10	2	0	31
4	1	0	2	2	10	9	3	0	27
5	0	0	3	3	14	8	3	0	31
6	0	0	3	2	7	12	2	0	26
7	0	0	0	1	2	1	0	0	4
8	0	0	0	1	0	0	0	0	1
total	2	1	12	16	49	47	10	0	137

32 pupils (i.e. 23% out of 137) perform equally good (or bad) in German and Turkish (grey fields), a difference of one level is found with 39 pupils (29%; from which 13 in favour of German, 26 in favour of Turkish). Out of these 71 children who approximate a balanced

bilingualism, 16 can be said to perform rather poorly (level 1 to 3, or level 3 in one language, level 4 in the other) – clearly children of special concern. Thus, balanced or nearly balanced bilingualism at a satisfying level (at least level 4 in both the languages) is found in 57 cases or 52% of the children. A clear dominance of German exists in 13 cases (10%), a clear dominance of Turkish in 53 cases (38%).

Irrespective of the dominance pattern, there could be a positive correlation between scores in German and Turkish so that relatively good competences in Turkish correspond to relatively good, though definitively weaker, even very much weaker competences in German. But this is not the case. The correlation coefficient of .065 (Spearman's rho; not significant) unambiguously indicates that such a correlation is non-existent.

For a researcher who, in principle, adheres to interdependence theory, (Cummins/Merrill 1986; Cummins 2000) this finding is a rather disappointing result. We have to conclude that up to preschool age, a large part of the Turkish-German children did not develop anything similar to interdependence between the competencies in first and second language.

IV. Towards the End of Grade 1

One year later the data collection was repeated with the same instruments in order to find out what progress had been made during the first year of schooling. This time 153 pupils were present on both the occasions.

Table 3: Changes in narrative competence up to the End of Grade 1

Second speech task – means of test scores in Turkish and German at different schools and difference in comparison with first speech task								
	school							**total**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
Turkish								
speech task 1	2,38	2,41	2,22	2,54	2,32	2,26	2,06	2,32
speech task 2	2,76	2,90	2,67	2,60	2,68	2,32	2,74	2,67
difference	+ 0,38	+ 0,49	+ 0,45	+ 0,06	+ 0,36	+ 0,06	+ 0,68	**+ 0,35**
German								
speech task 1	1,80	1,22	2,43	1,80	1,81	2,15	2,07	1,88
speech task 2	2,64	2,14	2,53	2,10	2,06	2,51	2,10	2,25
difference	+ 0,84	+ 0,92	+ 0,10	+ 0,30	+ 0,25	+ 0,36	+ 0,03	**+ 0,37**

The table shows that – not astonishingly – the children's performance has become better towards the end of grade 1 as compared with the time before entering school. However, in contrast to what could be expected, the level of improvement in both the languages is actually nearly the same. Even if we take into account that the initial position is different, insofar as Turkish starts from a higher level than German, we do not come to a very different conclusion. The rate of increase is about 15% in Turkish, about 20% in German. In either

case, the comparable improvement in both the languages means that the distance between first and second language has remained the same. Turkish still prevails over German by .42 points. The increase seems to be due to the effects of maturing, with equal influence on both languages. A definitive impact which the German school has on the development of children's German language ability, cannot be determined at this stage .

But this is not yet the whole story. As can be seen in the table, the differences between the two speech tasks illustrate the well-known statistical effect that starting from a low position raises the probability of increase and vice versa ("regression toward the mean"). School 7, for instance, which had very low Turkish scores in the first analysis, shows a huge increase in the mean of the test score, whereas the pupils of school 4 who started on a high level in Turkish, attain only a small improvement. Similarly, the increase in language proficiency in German is high in school 1 which started from a low level, but it is small in school 3, whose pupils already had good scores a year before. The standard deviation decreases from .58 to .47 in Turkish, and from .73 to .52 in German. We can say that by the end of grade 1, the language proficiency of pupils who had performed strongly and poorly has become somewhat more similar and that this holds true to a higher degree for German than for Turkish.

Accordingly, the table of language balances shows that the one-sided language relationships have been diminishing during the first school year:

Table 4: Language balances towards the end of grade 1

Second speech task – categorized levels in Turkish and German: language balances									
German	**Turkish**								**total**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	2	1	0	0	3
3	0	0	0	3	3	4	0	0	10
4	0	0	1	1	4	18	6	0	30
5	1	0	0	2	9	37	7	0	56
6	1	0	1	1	3	29	13	0	48
7	0	0	0	0	1	3	2	0	6
8	0	0	0	0	0	0	0	0	0
total	2	0	2	8	22	92	28	0	153

41 pupils (i. e. 27% of 153) attain equal levels in Turkish and German (fields with grey background) and therefore can be termed balanced bilinguals. A difference of one level only is found with 66 pupils (i. e. 43%, from them 9 cases in favour of German, 57 in favour of Turkish). Only 4 children can be said to be considerably weak in both the languages (level 4 in one language, level 3 in the other). That is to say that balanced or approximately balanced bilingualism on a more or less satisfying level (categories 4 to 7) can be found with 103 pupils who equal a percentage of 67%. This percentage is significantly higher than a year before when only 42% of the children could be characterized this way. A clear dominance of German appears with 5 children (3%), a clear dominance of Turkish with 41 (27%).

There is a weak correlation between Turkish and German language abilities with a coefficient of .193 (Spearman's rho, significant on the 5%-level).

At the same time, we find an intriguing difference in vocabulary development.

Table 5: Vocabulary increase in German

Second speech task 2 – verbal vocabulary in German – changes in comparison to first speech task									
category	Change in number of verbs	**number of children at schools 1-8**							**total**
		4	**2**	**3**	**4**	**5**	**6**	**7**	
1	**> +12**	9	6	0	6	0	6	3	30
2	**+8 to +12**	8	3	3	10	4	4	4	36
3	**+3 to +7**	0	2	7	6	5	2	8	30
4	**+2 to -2**	0	3	0	6	7	0	10	26
5	**-3 to -7**	0	0	2	2	2	1	2	9
6	**-8 to -12**	0	0	0	2	0	1	1	4
7	**< -12**	0	0	0	0	0	0	0	0
	total	17	14	12	32	18	14	28	135
	∅	+15,35	+12,93	+5,00	+5,97	+2,61	+10,14	+4,14	+7,39

While, as expected, the vocabulary in German shows an overall increase of more than 7 verbal types with few pupils performing poorer than a year before, the Turkish vocabulary shows an overall standstill, with changes in both directions – a development in need of further explanation:

Table 6: Vocabulary increase in Turkish

Second speech task– verbal vocabulary in Turkish – changes in comparison with first speech task									
category	Change in number of verbs	**number of children at schools 1-8**							**total**
		1	**2**	**3**	**4**	**5**	**6**	**7**	
1	**> +12**	0	1	0	0	4	2	2	9
2	**+8 to +12**	0	2	0	1	6	2	4	15
3	**+3 to +7**	3	4	3	5	3	1	6	25
4	**+2 to -2**	4	2	4	11	6	3	7	37
5	**-3 to -7**	5	3	3	8	0	3	6	28
6	**-8 to -12**	2	1	1	8	0	0	1	13
7	**< -12**	1	2	0	1	0	0	0	4
	total	15	15	11	34	19	11	26	131
	∅	-2,87	+ 0,13	-1,55	-3,44	+6,68	+3,64	+2,38	+0,41

Surprising as it may be, this result is in line with what Ibrahim Karasu (1995) has found in a research about 15 Turkish-German school beginners at Mannheim some years before. In his doctoral thesis he argues that entering the German school requires a certain amount of vocabulary, both general and instruction-related, whose acquisition is given priority by the children. This is particularly true, as this type of vocabulary partly belongs to a domain which might be more or less separated from family communication. He finds this effect stronger than average in the case of children who had not much contact with German before starting school and who had therefore developed a German vocabulary which was lacking in comparison with peers who had more contact with German before school.

To summarize, the children's ability to competently narrate in both languages has grown steadily during the first school year. In contrast to this observation, the increase lexical elements, clearly observable in the second language, is rather inconsistent and slim in the first language. It appears that we have to make a distinction between at least two types of progress in language acquisition, i.e. more general abilities on a macro-level, which seem to progress interdependently with age, and more specific knowledge which seems to be more dependent on external factors and therefore may progress differently in each of the languages.

V. In Grade 3

As already stated, the scores of the children's ability to competently narrate in grade 3 cannot be compared with the scores attained at the two previous periods because the children's speech task had changed. It would be likewise questionable to use vocabulary comparison as a means of demonstrating progress since the new speech task was clearly more extensive and dealt with another issue. Only at the structural grammatical level, would a longitudinal comparison seem to be justified. (But it has to be kept in mind that such measures can only be applied to one language). The result is that in both languages, progress in the individual's grammatical development is achieved, indicated by verbal suffixes in Turkish, by verb position in German.

Like the periods before, the children were asked to tell the story, this time the Mole's Story, in both the languages so that a comparison between performance in Turkish and in German remained feasible.

Table 7: Language dominance in Grade 3

Third speech task – means of test scores in Turkish and German at different schools								
	School 1	2	3	4	5	6	7	total
German	2,20	1,64	2,30	2,20	1,96	2,37	1,97	2,08
s	,31	,47	,41	,33	,34	,40	,34	,43
Turkish	2,57	1,93	2,10	2,03	2,32	2,69	1,97	2,18
s	,21	,46	,49	,35	,33	,47	,35	,46

The mean scores indicate that in the meantime the discrepancy has very much diminished; it remains a difference of only .1. Standard deviations also are rather narrow and they are very similar for Turkish and German. The obvious conclusion is that the children's ability to communicate in first and the second language had grown more parallel. This finding is corroborated by the table of the individual language balances:

Table 8: Language balances in grade 3

Second speech task – categorized levels in Turkish and German: language balance									
German	**Turkish**								**total**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	
1	0	0	0	0	0	0	0	0	0
2	0	1	0	1	1	0	0	0	3
3	0	0	4	6	2	2	0	0	14
4	0	0	2	19	23	5	0	0	49
5	0	1	1	16	22	17	1	0	58
6	0	0	1	1	10	14	0	0	26
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
total	0	2	8	43	58	38	1	0	150

In grade 3, 60 children (i. e. 40% of 150) attain the same level of performance in Turkish and German, with only 5 of them ranging in low categories in both languages. With 74 children (49%) we find a difference of no more than one level (28 cases in favour of German, 46 in favour of Turkish). The instances of clear dominance have significantly diminished, 4 in favour of German, 12 in favour of Turkish. The correlation between German and Turkish scores is high, coming to a coefficient of .411 (Spearman's rho, significant on the 1%-level).

VI. Conclusion

It comes as no surprise that entering the German school leads to the improvement of one's language proficiency in German in German. It has to be noted, however, that language abilities in Turkish also developed, initially even at the same speed. Only later on – as has to be concluded from the analysis of the third data collection – did they develop at a slower rate so that by grade 3 the pupils attain equal scores in both the languages.

It is more intriguing that correlation of the performance in either language was not found in the beginning, but emerged during the following years. It has to be concluded that interdependence of the languages in the bilingual individuals did not exist initially but came into being during a process of converging linguistic competences. In contrast to what could be assumed, this observation is not a tautology. Languages could be interdependent even if the mean score would be significantly different for each of the languages. The vice versa could also be true, in that, languages for bilingual individuals could independently develop from each other even if the mean scores within the group were very similar. Neither of these two conceivable possibilities is the case here. Instead, we find that the correlation coefficients are growing, while mean scores become more and more alike. It seems that transfer between languages does not occur if the language abilities are too differently developed, but is likely to take place in case of abilities developed in parallel.

Upon looking more carefully at the process, one can distinguish between the discourse level, represented in this case by the narrative competence, and the structural (or language

system) level, represented here by the use of vocabulary. It is clear from the progress from first to second data collection (and from the guesses we can make with regard to the third one), that catching up to the lead of the first language by the second one, as it were, occurs more slowly and more steadily on the discourse level, and more rapidly and with concentration on the weaker language, on the structural level.

It can be concluded that progress on the discourse level proceeds more from primary language socialization or "natural" linguistic development than from learning at school and that transfer occurs primarily on this level, though not under all circumstances. Interdependence manifests itself on one's level of discourse only if a certain critical mass of structural (or language system) elements are available. Discrepancy between expressive faculty, acquired in principle, and lack of vocabulary in one language seems to be a strong stimulus for acquiring the means of expression needed for one's purposes. In that way, discourse development can be said to determine or at least to influence the acquisition of structural elements. Since, in our case, the language of the family is clearly ahead of the language of environment (and school) at the beginning, we can attribute it an accelerating function for learning the second language, at least at this stage of development.

References

Cummins, Jim/Swain, Merrill (eds.) (1986): Bilingualism in education: Aspects of theory, research and practice. London u. A.: Longman.

Cummins, Jim (2000): Language, Power and Pedagogy. Bilingual Children in the Crossfire, Clevedon: Multilingual Matters.

Gogolin, Ingrid/Neumann, Ursula (eds.) (1997): Großstadt-Grundschule. Eine Fallstudie über sprachliche und kulturelle Pluralität als Bedingung der Grundschularbeit, Münster et al.: Waxmann.

Hepsöyler, Ender/Liebe-Harkort, Klaus (1988): Wörter und Begriffe - Lücken im Kindesalter = Verlust der Gleichberechtigung im Beruf und Gesellschaft. Auswertung eines Worttests bei türkischen Migrantenkindern in der Primarstufe - Vergleich mit nicht-migrierten Kindern: deutsche Schüler in der Bundesrepublik und türkische Schüler in der Türkei, Frankfurt a.M.: Lang.

Hepsöyler, Ender/Liebe-Harkort, Klaus (1991): Muttersprache und Zweitsprache. Türkische Schulanfängerinnen und Schulanfänger in der Migration - Ein Vergleich, Frankfurt a.M.: Lang.

Jeuk, Stefan (2003): Erste Schritte in der Zweitsprache Deutsch. Eine empirische Studie zum Zweitspracherwerb türkischer Migrantenkinder in Kindertageseinrichtungen, Freiburg im Breisgau: Fillibach.

Karasu, Ibrahim (1995): Bilinguale Wortschatzentwicklung türkischer Migrantenkinder vom Vor- bis ins Grundschulalter in der Bundesrepublik Deutschland. (= Werkstattreihe Deutsch als Fremdsprache, 50.) Frankfurt: P. Lang.

Pfaff, Carol W. (1991): Turkish in contact with German. Language maintenance and loss among immigrant children in Berlin (west), in: International Journal of the Sociology of Language, 97-129.

Preibusch, Wolfgang (1992): Die deutsch-türkischen Sprachenbalancen bei türkischen Berliner Grundschülern. Eine clusteranalytische Untersuchung, Frankfurt/M.: Lang.

Reich, Hans H. (2004a): Sprachstand und Sprachförderung bei zweisprachigen Kindern, in: Karakşoğlu, Yasemin/Lüddecke, Julian (eds.): Migrationsforschung und Interkulturelle Pädagogik. Aktuelle Entwicklungen in Theorie, Empirie und Praxis, Münster u. A.: Waxmann, 131-143.

Reich, Hans H. (2004b): Die sprachliche Entwicklung türkisch-deutscher Grundschüler in Hamburg. Projektbericht, Landau: Universität, Arbeitsbereich Interkulturelle Bildung (unpublished manuscript).

Reich, Hans H. (2005): Abschlussbericht der wissenschaftlichen Begleitung des Projekts „Sprachentwicklung zweisprachiger Kinder im Elementarbereich“, Landau: Universität, Arbeitsbereich Interkulturelle Bildung (unpublished manuscript).

Tracy, Rosemarie/Gawlitzek-Maiwald, Ira (2000): Bilingualismus in der frühen Kindheit. In: Hannelore Grimm (ed.): Sprachentwicklung, Göttingen et al.: Hogrefe, 495-535.

Bilingual Education – the German Experience and Debate[1]

Ingrid Gogolin

1. Introduction

The debate about the best ways to integrate children with an immigrant background in German schools, especially about adequate language education, has often been passionate, if not hot-headed – and very rarely based on empirical evidence. The title "the German experience and debate" on bilingual education is meant to indicate this state of affairs. Indeed we must admit that the German contribution to empirical research on bilingual education programs has until today been rather limited.

Why is this so? Why does German educational and linguistic research not play a significant role in the important and interesting evaluation of programs of language education for immigrant-minority children, be these bilingual programs or others? In order to be able to understand this, we have to understand education policies and practices implemented in response to the rising number of immigrant-minority children in Germany since the late 1960s. The first part of this contribution provides a brief introduction to the specific contextual factors which have to be considered in the German case. The second part gives an overview of recent research projects in Germany (see also the contribution by Hans H. Reich).

2. Language Education for Immigrant Children in Germany since the Late 1960s

As a consequence of the post-war immigration of workers and refugees, Germany eventually developed into a multilingual society. Today, the proportion of children growing up in families with an immigrant background is approximately one in four. As everywhere in the world, cities and urban areas are the most attractive destinations for immigrants. It is estimated that in German urban areas the proportion of immigrant-minority children in schools is around one third. At least 100 immigrant-minority languages are spoken by chil-

[1] This text outlines a general argument that will be further elaborated in an article to be published in 2006.

dren with an immigrant background who attend German schools (see Fürstenau/Gogolin/Yagmur 2003; Chlosta/Ostermann/Schroeder 2003).

Why do we have to estimate, why do we not know precisely how many children in German schools have an immigrant background or how many languages are spoken on German territory? The reasons can be found in the circumstances which are also partly responsible for the fact that the contribution of German research has been rather limited with regard to questions of language education for children with an immigrant background who grow up and live with two or more languages. Bilingual school programs are an element of this broader question: They are a very specific approach to giving them access to literacy.

As a matter of fact, hardly any reliable official statistics are available, be it on the ethnic-cultural composition of Germany's population, be it on the socio-linguistic situation in Germany. There is no reliable information either about the number of speakers of languages other than German or the number of languages spoken by them. The absence of such data is due to the traditional self-conception of the German nation-state. Until recently and despite all evidence, Germany considered itself to be a non-immigration country and by implication a monolingual country. Thus, there are only official statistics about foreign nationals ('Ausländer') living in Germany permanently or temporarily.[2] As everybody knows, such data are unsuitable for the assessment of cultural and linguistic diversity. First (and this is trivial), the equation of 'nation state' and 'language' would be inadequate. Second, a growing number of immigrants in Germany possess German citizenship. They may have the status of 'Aussiedler', i.e. ethnic German immigrants from Eastern Europe; or they may be naturalised immigrants from anywhere else who applied for German citizenship after having lived in Germany for at least eight years; or they may be children of binational couples with one parent carrying a German passport. The absence of adequate statistical data about the socio-linguistic, ethnic and cultural composition of the German population after more than 50 years of immigration is indeed indicative of the German integration policy as a whole and especially of language education policies for immigrant-minority children. It highlights the fact that educational policies were based on the dictum 'Deutschland ist kein Einwanderungsland' – Germany is no immigration country. Thus, the country did not develop a systematic integration policy but based its integrative efforts on its aliens laws. This principle also affected policies in the field of education as the legal status of immigrant children was – and in some regional states still is – decisive for eligibility for different types of educational measures aimed at supporting immigrant children's academic advancement in school.[3]

---

[2] This is indeed an element of the unsatisfying statistical situation regarding the whole problem of immigration in Germany; see for the criticism of this situation for example: OECD 2005 and BMBF 2005.

[3] Entorf/Minoiu (2004) show that immigration law is certainly not irrelevant for school performance of immigrant students, although other factors such as social class and command of the official school language play the most important role.

Legislation introduced from the 1960s onwards (in fact at the time only in the 'old' Federal Republic of Germany) aimed at opening up the school system to immigrant-minority children and was accompanied by a discourse of equal opportunities. Measures taken to put this rhetoric into practice included the introduction of language courses entitled 'Deutsch als Zweitsprache' – German as a Second Language. These have been taught in so-called reception classes or separate courses for new immigrants. The support of new immigrants in classes of this type can last between six months and a year, in some cases up to two years. In the meantime, some Länder have introduced flexible models of transferring students from these classes to regular school education (see for example the concept of 'Deutsch als Zweitsprache' in the Land Saxony/Sachsen). But German as a Second Language was not – and is not until today – made part of the teaching of German in the regular framework of the school system (see Reich 2000). In this respect, the situation in Germany differs from that in some other European countries.

As we know from recent research, roughly 75 to 80% of the immigrant minority population in German schools today were born in Germany or have grown up in the country. In contrast to newly arrived immigrant children these children have only very little or no chances to benefit from a specific language education or other measures addressing their special needs. They may temporarily receive what is called 'Förderunterricht' – special tuition. Schools with a large share of immigrant children normally obtain extra support in order to enable them to better deal with this situation. The measures differ in the Länder, but in most cases schools are provided with extra teaching hours for this purpose. There has hardly been any systematic development of these courses; teacher training for this special task has hardly been available; and there has been very little monitoring of what is actually being done with these resources in schools. It is reported that in many cases this extra tuition takes place more or less independent from the regular school day (Kuhs 2000). Until recently, the appropriateness and effectiveness of these kinds of measures was not evaluated.[4] Some of the Länder are now introducing different procedures according to which schools can receive extra resources if they present explicit programs to the authority, including a procedure of evaluation and verification of the success of their efforts (e.g. in Northrhine-Westphalia since 2004, in Hamburg since 2005).

Another characteristic of language education measures introduced in the 1960s was the teaching of immigrant minority languages in supplementary lessons. Since 1964, 'foreign children' in West Germany were in principle guaranteed the same educational opportunities as German children. If it was considered beneficial for this aim, teaching of their languages of origin (officially called 'mother tongue-teaching') could be provided in addition to the

[4] A lack of evaluation studies is no special or excepional, but a typical characteristic feature of the German school system: there is no tradition of controlling the adequateness or effectiveness of investments of any kind, be it financial or conceptual, with regard to the 'outcomes' in terms of student achievement. The discourse about school effectiveness in the sense of output-oriented regulation of the German school-system is a concomitant of the transformation of a public 'Bildungssystem' into an education system with stronger links to the economicy. For the controversy about this development see for example Lohmann 2002.

regular curriculum. The so-called 'mother tongue teaching' should, according to the relevant decrees, contribute to the social integration of the students „for the duration of their stay in the Federal Republic of Germany" while at the same time „preserving their linguistic and cultural identity". The underlying agenda of these recommendations was, similar to the policies in other European immigration countries and the official European Community policy in those days, a 'rotation-perspective'.

The organisational measures the Länder could opt for were:

- Supplementary teaching of the native language as a voluntary option for immigrant children attending mainstream classes. This teaching could take place in addition to the regular curriculum and school day and was not to exceed five lessons per week.
- 'Mother tongue teaching' in place of the first or second obligatory foreign language (usually English or French). This special measure for 'foreign students' was intended to 'protect' them from the burden of learning further foreign languages when their command of the German language was poor. In practice, accepting this offer could mean a severe limitation of the potential school success, as in Germany the possibility of receiving the highest qualification (Abitur) is bound to a fixed catalogue of obligatory foreign languages; none of the immigrant-minority languages belongs to this catalogue. Therefore a 'foreign student' who accepted this offer was either excluded from paths leading to the highest qualification or had to invest extra time and energy in order to learn one of the obligatory foreign languages and be admitted to the Abitur-exams. In practice this often meant that for students who opted for this kind of mother tongue education during their regular school careers, the only way to obtain the Abitur was via adult education courses.
- 'Mother tongue' as a subject and as language of instruction in reception classes for pupils of the same nationality. This type of teaching was established for the large numbers of migrants of the same origin who came to Germany in the context of 'guest worker'-recruitment. Meanwhile, this model has disappeared due to the changes of immigration patterns.

Within this framework the responsibility for organisation, financing and controlling as well as curriculum development of 'mother tongue-teaching' followed two different patterns: Around half of the Länder assumed the responsibility for this teaching; the other half invited the countries of origin to take on this responsibility, including the recruitment of teachers and the provision of curricula and textbooks. Within the latter model, the respective Land contributed to the costs, e.g. by providing classrooms free of charge. There was and still is a wide variety of legal and organisational patterns of 'mother tongue-teaching' in the different Länder,[5] but it was only temporarily and in exceptional cases considered to be a regular and integrated part of the school system, and it was a continuous issue of political

[5] In the meantime, 'mother tongue-teaching' is already abolished or a phaseout model in most of the Länder.

debates (see summaries of these debates in the late 1990s: Reich 2000a; 2000b).[6] The characteristics of this instruction are discontinuity rather than continuity, isolation from the teaching in 'regular schools' rather than co-operation or co-ordination, non-systematic development and use of curricula or teaching material rather than systematic approaches, and last but not least: hardly any comprehensive qualification of teachers. Participation in 'mother tongue teaching' is voluntary[7], and most of it takes place in extra school hours. It is not surprising that statistical data about participation in these courses are patchy. The participation rate is different from language group to language group, from region to region; probably only a small fraction of immigrant-minority children have access to or make use of these offers (figures will be provided in the final version of this text as far as they are available).

To summarize this brief description: the actual 'mother tongue teaching' in Germany can certainly not be characterized as a systematic contribution to bilingual development, let alone bilingual education. As a consequence of the ways in which it was implemented in the German education system, the mere existence of this measure cannot serve as an indicator in favour or against any hypotheses about the effects of (language) education on immigrant-minority children in Germany.

3. Research on bilingual education programs

As a matter of fact, only a handful of 'bilingual education'-programs has ever been introduced in Germany. They are considered a very specific approach in the broader framework of the development of strategies for learning and teaching in linguistically and culturally plural classrooms. Experts in the field emphasise the necessity of developing and evaluating a range of different strategies with respect to specific features of immigration to Germany (see Gogolin/Neumann/Roth 2003: 30ff).

Some 'bilingual education'-programs have in recent years been developed as experimental projects. They include bilingual primary schools where two languages are taught right from the beginning. While bilingual classes in secondary schools are already quite common – mostly with French or English as the second language - only six of the sixteen Lander offer classes in primary schools where children are able to learn two languages from first grade onwards. All in all, there are no more than two dozens of these experimental projects

[6] It seems to be a standing rule in these debates to characterize the opponent's arguments as 'ideological'. This pattern not only appears in political debates, but also in so called scientific discourse.

[7] With one historical exception in the Land Hessen; see J. Schroeder 2001

which have immigrant-minority languages as the partner lanugage.[8] Even fewer of these projects have been studied or evaluated.

Only a minority of bilingual schools or classes deal with the problem of immigrant-minorities. The largest share of models addresses students from autochthonous minorities (e.g. German-Danish schools in Schleswig-Holstein; German-Sorbian schools in the Länder Brandenburg and Saxony). Another version is the 'classic' model of introduction of a foreign language as medium of instruction to the curriculum. This kind of 'bilingual education' is mainly practised with English as second language, in some cases also with French. Another version of 'bilingual education' is geared to special groups of immigrants such as children of members of the diplomatic corps or of managers of international companies.

Most of the experimental programs dealing with immigrant-minority languages have been designed following US-American or Canadian examples. In theory, half of the children in a class are meant to have a bilingual family background; the other half is meant to be monolingual in German. Classes are held in both languages, but there are differences in the organization of the actual teaching. Two models from Berlin and Hamburg illustrate the varieties:

In the Berlin model the children of a class are assigned to two different groups: the bilinguals and the monolinguals. Each group receives their early literacy education in their mother tongues and is introduced to the other language (the so-called partner language) separately. Some subjects are taught in joint groups in the non-German partner language, whereas mathematics is taught jointly in German. In this model, both languages are taught in an equal number of lessons. The teaching staff is made up of an equal number of native speakers of the two languages. The Hamburg model is different. The teachers of the non-German partner languages are civil servants of their respective countries of origin. They offer twelve additional lessons. Thus the two languages are not accorded equal time and relevance, rather German dominates. The teachers are free to divide the classes in language groups or to have the children work together as teams in the classrooms. Access to literacy is provided in both languages, either in parallel or with a delay of only a few weeks.. With regard to the partner language the children are not separated into language groups but receive language tuition together. The school reports assess the children's performance in the partner language with one mark, and five marks are given for the different subjects that are taught in German.

Only very few of these experimental programs have been or continue to be the object of empirical research. A more fully developed version of this text will contain a detailed description and discussion of research results, including results of an experimental model in

[8] The final version of this text will give the accurate number of bilingual primary schools in Germany and the term 'bilingual education' with its wide range of meanings will be dicussed more thoroughly.

Berlin which was carried out in the 1980s (Nehr et al. 1988). The following brief description only addresses themes, methods and results of current research projects.[9]

3.1 The German-Italian school in Wolfsburg

The German-Italian school in Wolfsburg was one of the very first experimental bilingual school programs in Germany. This model was observed and accompanied by researchers for a number of years. The observation covered the four years of primary school. A final evaluation compared school reports and the results of achievement tests in order to assess the students` competency in reading, orthography and mathematics. The authors stress the methodological limits of their evaluation. Given the methods available they had to use instruments which were calibrated for monolingual children. For the assessment of Italian a new instrument had to be developed.

The main results of this study are the following:

- Monolingual German children received better marks in German than bilingual children, regardless of whether one or both parents of these children were Italian.
- On the other hand, bilingual German-Italian children received better marks in Italian.
- The reading test confirmed the tendencies which were visible in the school reports: 81% of the German, 64% of the Italian and 72% of the German-Italian children turned out to be „fluent readers" (the norm being 65.9%). However, 25% of the bilingual children turned out to require special support in German reading and orthography.
- In mathematics the monolingual German children also seemed to benefit more from the program as indicated by their marks in mathematics. The achievement test however suggested less pronounced differences.

As far as academic performance in school is concerned, the Wolfsburg experiment indicates that monolingual German children seem to profit more from bilingual teaching than bilingual children. Apart from academic performance, the social atmosphere in the school was studied as well. In this respect the results were remarkably positive. A very positive school climate is one of the merits of the bilingual program. Students considered teachers as being positive and supportive; feelings of integration and good relations to fellow students were underlined. The overall satisfaction with the situation at school was higher among the children with a minority background than among the German children.

The authors of the study stress that monolingual children were not disadvantaged in the Wolfsburg bilingual school. The remaining difference in performance between the monolingual children and those with an immigrant background was in their view due to differences in social status on one hand and to insufficient support for children of lower performance on the other. But these explanations are only hypotheses, as both aspects were not part of the research design.

---

[9] For a detailed description see also Neumann/Roth 2005.

3.2 The German-Spanish class in Nuremberg

Another empirical research project studied third graders in a German-Spanish bilingual class in Nuremberg. The children were given a creative writing exercise, which was then evaluated in terms of orthography, syntax and vocabulary (Kupfer-Schreiner 1994).

The author did not find a uniform process of learning German. Rather, the process was influenced by prior language proficiency. The bilingual Spanish-German group showed the most noticeable improvement in learning German. The study also included children with other native languages who had not received bilingual tuition at school. These children showed the lowest improvement in German language skills. However, as they had already entered the course with a very high level of proficiency, they achieved the best results compared to the class average.

With regard to Spanish, the author found an enormous improvement of orthography in the bilingual group, which she refers to as a "parallel development" of languages. She was able to show that the learning process did not follow a linear line of progress for all students, but that progress took place in stages. The author stated that the teaching did not necessarily have a direct and immediate effect, but effects might show in later stages of language development. Kupfer-Schreiner thus emphasized the possible long-term effects of bilingual education programs. To conclude, the author highlighted the indisputable improvement of the children's general linguistic skills and their increased capacity to use both languages creatively.

3.3 The „Europa-Schulen" Berlin (SESB)

The Berlin model "Europa Schulen" was already mentioned. It was only very tentatively evaluated through a critical analysis of the experimental design (Zydatiß 2000) and a pilot study with some empirical data, focusing on the language development in an Italian-German class (Graefe-Bentzien 2001). The pilot study found that, in the bilingual class, the simultaneous acquisition of two languages does not interfere with competence in the first language. On average, the experimental group reached the same competence level as the respective monolingual control group. Children with a bilingual background made steady progress in learning German, but the progress they made in Italian remained unsatisfactory. The author emphasized the preliminary character of her findings, given the fact that the children tested had not yet finished primary school. Pre-school institutions and family background were identified as additional explanatory factors that contribute to the learning progress in both languages; they proved to influence the overall attitude towards communication as well as the development of lexical concepts and achievement in other school subjects.

3.4 The "Bilingual Primary School"-experiment in Hamburg

The last example concerns the evaluation of the Hamburg experiment. The evaluation included bilingual classes for Italian, Spanish and Portuguese students and was carried out by my colleagues Ursula Neumann, Hans-Joachim Roth and myself. Our research design consisted of the assessment of oral language development and literacy in both languages. Relevant context factors such as socio-economic background and parental attitudes towards language education were also analyzed. Another aspect of the research was the observation of the actual teaching process in classrooms. External performance control tests (following PIRLS, Bos et al. 2004) were performed on fourth graders.[10] The aim of this evaluation was to assess the contribution of teaching methods to the development of bilingual skills, not only to the development of L2.

A range of instruments was developed for the continuous observation of progress in both languages and tested within the framework of the evaluation. Oral speech was evaluated by using a procedure developed by Hans H. Reich which follows Harald Clahsen's profile analysis. Originally developed for German, the instrument was adapted to Italian, Spanish and Portuguese. Tests which take into account the different grapheme systems were also developed for the evaluation of writing skills in both languages. Some of our preliminary findings are as follows:

- The bilingual education program in our model is carried out in multilingual, and not just bilingual, classes. Apart from the four languages which were formally included in the project there were several other native languages present. In many families more than two languages are used. Hence, the experiment mirrors the multilingual character of the student population in Hamburg.
- Although the actual composition of the classes did not fulfil the distributional criteria of the experimental model (which was 50:50 German monolingual and bilingual children), the development of reading comprehension was more than satisfactory. With very few exceptions, all children achieved a high score in the test we used.
- In the group of second language learners, the acquisition of reading comprehension in German also proceeded well. After two years of learning, the proficiency level of this group was no longer significantly different from that of other students.
- In third grade, reading comprehension in the partner languages was still highly dependent on preschool linguistic attainments. Children who were bilingual when they started primary school seemed to develop a balanced bilingualism with regard to reading comprehension, whereas children whose dominant language was German remained dominant in this language.

The proficiency level monolingual German children achieved in the partner languages appeared to be highly dependent on factors other than linguistic preconditions. For instance,

[10] See Gogolin/Neumann/Roth 2001, 2003; Roth 2002, 2003; Hansen 2001; Owen-Ortega 2003.

the ratio of boy and girls within a class seems to have an indirect effect on achievement: Apparently, children of this age prefer to communicate with peers of their own gender. Where few peers of the same gender attended one class, this appears to have limited the effects of the formal teaching of the second language.

4. Conclusion

The findings of existing evaluation studies in Germany show that there is no clear empirical evidence for either position in the controversy about the effects of bilingual school programs. There is only a small number of recent studies, and these partly suffer from empirical weaknesses in the research design. Nevertheless these studies – as well as earlier research carried out in other European countries - are clearly supportive of bilingual education in general (although not necessarily of bilingual school programs). The impact of supporting the first language on the acquisition of L2 is neutral in the worst case. There is no evidence of negative effects on the acquisition of a second language (Felix 1993). The question as to whether bilingual children profit from bilingual education can therefore be answered with 'yes', as there are no negative effects on L2-development, but unambiguous positive effects on bilingual development. What remains unclear, however, is the question of whether they would benefit equally (or perhaps even more) from other measures, if only L2 were concerned. Answering this question would require comparative experimental settings and research, which has been lacking up to now.

What emerges clearly however, is the tremendous failure of the large-scale experiment we carried out in Germany over the last four decades, i.e. treating immigrant-minority children with a German-only approach, in a small number of cases accompanied by attempts to teach their L1 under conditions which do not meet generally accepted quality criteria. Studies such as TIMS, PISA or PIRLS, to name just a few, show the results of this approach: Students with an immigrant background tend to score well below average. These studies clearly indicate that immigrant minority pupils in Germany do not really learn German in German schools – at least they do not acquire literacy to an extent that is sufficient for satisfactory academic achievement and a successful school career. In contrast, recent studies (one of them not yet published) indicate that for immigrant-minority children the real hurdle lies in learning German, and not in language learning as such. A study in which all fourth graders in Hamburg were tested with instruments similar to the PIRLS-instruments ('KESS4') shows that while the achievement of immigrant-minority children was considerably below the average of non-immigrant children in all areas tested in German it was not the case in the tests of English.

Considering these results, I would like to draw attention to the question whether better achievement in the majority language can be – or should be – the most relevant concern in debates about language education for immigrant-minority children. Is it not inappropriate

to ask about positive effects on only one of the languages concerned, namely German, instead of looking at the effects on both languages, i.e. German and the first language? To rephrase the question more radically: Is it not inappropriate to focus only on effects on L2-acquisition instead of searching for the best ways to teach children who live in two languages, taking into account their specific living conditions? If future research takes this concern seriously, it will have to acknowledge that high proficiency in the majority language and the development of bilingual competences are by no means alternatives, but the complementary sides of the same coin. Indeed, it is time to end ideological debates and accept, with a more pragmatic attitude, that we have to find best solutions given the irreversible fact that we live in multilingual societies.

References

BMBF (Bundesministerium für Bildung und Forschung) (ed.) (2005): Bildungsdaten und Migrationshintergrund: Wege zur Verbesserung der amtlichen Statistik. Schriftenreihe Bildungsreform, Nr. 14. Bonn/Berlin.

Bot, Kees de/Driessen, Geert /Jungbluth, Paul (1989): De effectiviteit van het onderwijs en eigen taal en cultur. Prestaties van Marokkanse, Spaanse en Turkse leerlingen. Nijmegen: Institut voor Toegepaste Sociale Wetenschappen/Institut voor Toegepaste Taalkunde.

Bot, Kees de/Driessen, Geert /Jungbluth, Paul (1991): An evaluation of migrant teaching in the Netherlands. In Jaspaert, Koen/Kroon, Sjaak (eds.): Ethnic minority languages and education. Amsterdam: Swets & Zeitlinger, 25-123.

Bühler-Otten, Sabine/Fürstenau, Sara (2004): Multilingualism in Hamburg. In: Extra, Guus/ Yağmur, Kutlay (ed.): Urban Multilingualism in Europe. Immigrant Minority Languages at Home and School. Clevedon, Buffalo, Toronto: Multilingual Matters Ltd., 163-191.

Chlosta, Christoph/Ostermann, Torsten/Schroeder, Christoph (eds.) (2003): Die „Durchschnittsschule" und ihre Sprachen. Ergebnisse des Projekts „Sprachenerhebung an Essener Grundschulen". In: ELISE 1; http://www-elise.uni-essen.de; siehe auch: http://www.uni-essen.de/daz-daf/Projekte/index/spreeg.htm (Stand: 19.04.2005).

Entorf, Horst/ Minoiu, Nicoleta (2004): PISA Results: What a Difference Immigration Law Makes. Bonn: IZA Discussion paper No. 1021.

Felix, Sascha (1993): Psycholinguistische Untersuchungen zur zweisprachigen Alphabetisierung. Gutachten im Auftrag der Berliner Senatsverwaltung für Schule, Berufsbildung und Sport. Passau: Lehrstuhl für Allgemeine Linguistik der Universität Passau (unpublished manuscript).

Fürstenau, Sara/Gogolin, Ingrid/Yağmur, Kutlay (eds.) (2003): Mehrsprachigkeit in Hamburg. Ergebnisse einer Sprachenerhebung an den Grundschulen. Münster/New York: Waxmann.

Gogolin, Ingrid (1994): Der monolinguale Habitus der multilingualen Schule. Münster et al..: Waxmann.

Gogolin, Ingrid/Neumann, Ursula/Roth, Hans-Joachim (2001): Auswertung der ersten Sprachstandserhebung der portugiesisch-deutschen Klasse, Schuljahr 2000/01. Hamburg: Univ. Hamburg (unpublished manuscript).

Gogolin, Ingrid/Neumann, Ursula/Roth, Hans-Joachim (2003): Bericht 2003. Schulversuch Bilinguale Grundschulklassen in Hamburg. Hamburg: Univ. Hamburg, Arbeitsstelle Interkulturelle Bildung (unpublished manuscript).

Graefe-Bentzien, Ulrike (2001): Evaluierung bilingualer Kompetenz. Eine Pilotstudie zur Entwicklung der deutschen und italienischen Sprachfähigkeiten in der Primarstufe beim Schulversuch der Staatlichen Europa-Schulen Berlin (SESB) – Evaluation of bilingual competence. Dissertation an der Freien Universität Berlin, online-Veröffentlichung: http:\\www.diss.fu-berlin.de/2001/14/index.htmlHansen, Christiane (2001): Bilingualer Schriftspracherwerb am Beispiel einer italienisch-deutschen Modellklasse einer Hamburger Grundschule. Hausarbeit im Rahmen der Ersten Staatsprüfung. Hamburg.

Kupfer-Schreiner, Claudia (1994): Sprachdidaktik und Sprachentwicklung im Rahmen interkultureller Erziehung. Das Nürnberger Modell. Ein Beitrag gegen Rassismus und Ausländerfeindlichkeit. Weinheim: Dt. Studienverlag.

Lohmann, Ingrid (2002): Bildungspläne der Marktideologen – Ein Zwischenbericht. In: Vierteljahresschrift für die wissenschaftliche Pädagogik 3, 267-279.

Nehr, Monika [et al.] (1988): In zwei Sprachen lesen lernen – geht denn das? Erfahrungsbericht über die zweisprachige Alphabetisierung. Weinheim u. Basel: Beltz.

Neumann, Ursula & Roth, Hans-Joachim (2004 in print): Bilinguale Grundschulklassen in Hamburg – Ein Werkstattbericht. In: Grenzgänge 11 (2004), H. 21, 29-56.

OECD (2005): Trends in International Migration: SOPEMI, 2004 edition. Paris (OECD).

Owen-Ortega, Julia (2003): Schriftspracherwerb bilingualer Kinder am Beispiel einer spanisch-deutschen Klasse unter besonderer Berücksichtigung ihrer Strategien des Orthographieerwerbs in der Alphabetisierungsphase. Hausarbeit im Rahmen der Ersten Staatsprüfung. Hamburg.

Ramirez, David J./ Yuen, Sandra D./ Ramay, Dena R. et al. (1991) Final Report: Longitudinal study of structured English immersion strategy, early-exit and late-exit transitional bilingual education programs for language-minority children. Report submitted to the US-Department of Education. San Mateo, CA: Aguirre International.

Reich, Hans H. (2000): Deutsch: Sprache. In: Reich, Hans H./ Holzbrecher, A./ Roth, Hans-Joachim (eds.): Fachdidaktik interkulturell. Ein Handbuch. Opladen: Leske + Budrich, 235-256.

Reich, Hans H. (2000a): Machtverhältnisse und pädagogische Kultur. Die Legitimierung des Unterrichts in den Herkunftssprachen von Migranten als Gegenstand eines internationalen Vergleichs. In: Gogolin, Ingrid/ Nauck, Bernhard (eds.): Migration, gesellschaftliche Differenzierung und Bildung. Oplade: Leske + Budrich, 343-364.

Reich, Hans H. (2000b): Die Gegner des Herkunftssprachenunterrichts und ihre Argumente. In: Deutsch Lernen, Heft 2, 112 -125.

Reich, Hans H./ Roth, Hans-Joachim [et al.] (2002): Spracherwerb zweisprachig aufwachsender Kinder und Jugendlicher. Ein Überblick über den Stand der nationalen und internationalen Forschung. Hamburg: Behörde für Bildung u. Sport.

Roth, Hans-Joachim (2002): Il gatto va sull´albero – va sull´albero il gatto. Satzmuster und Sprachstand italienisch-deutscher Schulanfänger (2002). Hamburg: Univ. Hamburg, Arbeitsstelle Interkulturelle Bildung (unpublished manuscript).

Sandfuchs, Uwe/Zumhasch, Clemens (2002): Wissenschaftliche Begleituntersuchung zum Schulversuch Deutsch-Italienische Grundschule Wolfsburg – Reflexionen und ausgewählte Ergebnisse. In: interkulturell, Heft 1/2, 104-139 .

Schroeder, Joachim (2001): Länderbericht: Hessen. In: Gogolin, Ingrid/ Neumann, Ursula/ Reuter, Lutz (eds.): Schulbildung für Kinder aus Minderheiten in Deutschland 1989 – 1999. Münster: Waxmann-Verlag, 187-205.

Skutnabb-Kangas, Tove (1986): Multilingualism and the education of minority children. In: Skutnabb-Kangas, T. & Cummins J. (Ed.): Minority Education: From Shame to Struggle. Clevedon, Philadelphia (Multilingual Matters)

Sächsisches Staatsinstitut für Bildung und Schulentwicklung (o.J./2004): Lehrplan Deutsch als Zweitsprache. URL der Ressource:
http://www.sn.schule.de/~ci/1024/lp_abs_landesliste_gs.html#Dgs

Verhoven, Ludo (1994): Transfer in bilingual development: The linguistic interdependence hypothesis revisited. In: Language Learning 44/3; 381-415.

Westerbeek, Karin/ Wolfgram, Peter (1999): Deltaplan en het tij. 7 jaar taalbeleid in Rotterdam: Deltaplan Taalbeleid Primair Onderwijs. Rotterdam (Het Projectbureau/CED)

Wode, Henning (2004): Frühes Fremdsprachenlernen. Englisch ab Kita und Grundschule: Warum? Wie? Was bringt es? Kiel: Verein für frühe Mehrsprachigkeit an Kindertageseinrichtungen und Schulen FMKS e.V.

Zydatiß, Wolfgang (2000): Bilingualer Unterricht in der Grundschule. Entwurf eines Spracherwerbskonzepts für zweisprachige Immersionsprogramme. Ismaning: Hueber.

List of Contributors

Monica Axelsson is senior researcher at the Centre for Bilingual Research, University of Stockholm, Sweden. (monica.axelsson@biling.su.se)

Alan Cheung is a researcher at the Success for All Foundation, USA. (acheung@SuccessForAll.net)

Geert Driessen is senior researcher at the Institute for Applied Social Sciences (ITS), Radboud University Nijmegen, The Netherlands. (G.Driessen@its.ru.nl)

Ingrid Gogolin is Professor of Education at the Institute for Comparative and Multicultural Studies, University of Hamburg, Germany. (gogolin@erzwiss.uni-hamburg.de

Julia Kuder is a candidate for a Master's degree in Mathematics in the Department of Mathematics and Statistics, Boston University, USA. (JuliaFK@bu.edu)

Hans H. Reich is Professor of German as a Second Language and Intercultural Pedagogy at the Institute for Intercultural Education, University of Landau-Koblenz, Germany. (iku@uni-landau.de)

Christine Rossell is Professor of Political Science in the Political Science Department, University of Boston (USA). (crossell@bu.edu)

Robert Slavin is Co-Director of the Center for Research on the Education of Students Placed at Risk, Johns Hopkins University, USA, and Chairman of the Success for All Foundation. (rslavin@successforall.net)

Programme on Intercultural Conflicts and Societal Integration

The Programme (AKI) at the Social Science Research Center Berlin (WZB) focuses on the synthesis of research results from different disciplines in the thematic field of immigrant integration and intercultural conflicts. It thus aims to contribute to discussions about future directions of academic research and to provide accessible and sound evaluations of existing knowledge and policy options.

The underlying assumption of the Programme is that a wealth of models and findings are available in academic scholarship that could help German society deal with challenges arising from migration and ethnic plurality. However, this potential is often not fully exploited. The Programme aims to address this deficit by helping to promote co-operation and communication between academics, policy-makers and the wider public. It also aims to encourage interdisciplinary dialogue and to contribute to a higher profile within academia and German society of research into migration and intercultural conflicts.

The Programme began in 2003 and is funded by the Federal Ministry of Education and Research and affiliated with the research area Civil Society, Conflict and Democracy. AKI has a steering group as well as its own advisory committee comprising experts with a policy or media background and scholars from Germany and abroad.

Topics include

- migration and illegality
- language acquisition, educational participation and intergenerational processes of integration
- cultural differences, social identities and educational achievements: research on stereotypes and discrimination
- urban segregation and interethnic conflicts
- official data on the integration of individuals with an immigrant background and of members of ethnic minorities

Publications include a newsletter which is available in an electronic and a printed version.

AKI: Dr. habil. Karen Schönwälder (head), Dipl.-Soz. Janina Söhn (researcher), Manuela Ludwig (secretariat)

Members of the steering group: Prof. Dr. Klaus J. Bade, Osnabrück, Prof. Dr. Hartmut Esser, Mannheim, Prof. Dr. Wilhelm Heitmeyer, Bielefeld, Prof. Dr. Amélie Mummendey, Jena, Prof. Dr. Friedhelm Neidhardt, Berlin.

Arbeitsstelle Interkulturelle Konflikte und gesellschaftliche Integration (AKI)
Wissenschaftszentrum Berlin für Sozialforschung (WZB)
Reichpietschufer 50
10785 Berlin
Tel.: 030-25491-352
aki@wz-berlin.de
www.aki.wz-berlin.de